MAECO-408

Statistics

MA ECONOMICS

2nd Semester

Rajiv Gandhi University

www.ide.rgu.ac.in

Statistics

MA [Economics]

Second Semester

MAECO 408

RAJIV GANDHI UNIVERSITY

Arunachal Pradesh, INDIA - 791 112

BOARD OF STUDIES		
1.	**Prof. N. C. Roy** Dept. of Economics, RGU	Chairman (Ex-officio)
	Prof. S. K Nayak Dept. of Economics, RGU	Member
2.	**Dr. Lijum Nochi** Dept. of Economics, RGU	Member
	Dr. Maila Lama Dept. of Economics, RGU	Member
3.	**Prof. (Ms.) V. Upadhyay** Dept. of Economics, RGU	Member Secretary

Authors:

V.K. Khanna,
S.K. Bhambri,
S. Kalavathy,
J.S. Chandan,
C.R. Kothari,
GS. Monga,
Vikas Publishing House,

Vikas revised edition, 2021

About the University

Rajiv Gandhi University (formerly Arunachal University) is a premier institution for higher education in the state of Arunachal Pradesh and has completed twenty-five years of its existence. Late Smt. Indira Gandhi, the then Prime Minister of India, laid the foundation stone of the university on 4th February, 1984 at Rono Hills, where the present campus is located.

Ever since its inception, the university has been trying to achieve excellence and fulfill the objectives as envisaged in the University Act. The university received academic recognition under Section 2(f) from the University Grants Commission on 28th March, 1985 and started functioning from 1st April, 1985. It got financial recognition under section 12-B of the UGC on 25th March, 1994. Since then Rajiv Gandhi University, (then Arunachal University) has carved a niche for itself in the educational scenario of the country following its selection as a University with potential for excellence by a high-level expert committee of the University Grants Commission from among universities in India.

The University was converted into a Central University with effect from 9th April, 2007 as per notification of the Ministry of Human Resource Development, Government of India.

The University is located atop Rono Hills on a picturesque tableland of 302 acres overlooking the river Dikrong. It is 6.5 km from the National Highway 52-A and 25 km from Itanagar, the State capital. The campus is linked with the National Highway by the Dikrong bridge.

The teaching and research programmes of the University are designed with a view to play a positive role in the socio-economic and cultural development of the State. The University offers Undergraduate, Post-graduate, M.Phil and Ph.D. programmes. The Department of Education also offers the B.Ed. programme.

There are fifteen colleges affiliated to the University. The University has been extending educational facilities to students from the neighbouring states, particularly Assam. The strength of students in different departments of the University and in affiliated colleges has been steadily increasing.

The faculty members have been actively engaged in research activities with financial support from UGC and other funding agencies. Since inception, a number of proposals on research projects have been sanctioned by various funding agencies to the University. Various departments have organized numerous seminars, workshops and conferences. Many faculty members have participated in national and international conferences and seminars held within the country and abroad. Eminent scholars and distinguished personalities have visited the University and delivered lectures on various disciplines.

The academic year 2000-2001 was a year of consolidation for the University. The switch over from the annual to the semester system took off smoothly and the performance of the students registered a marked improvement. Various syllabi designed by Boards of Post-graduate Studies (BPGS) have been implemented. VSAT facility installed by the ERNET India, New Delhi under the UGC-Infonet program, provides Internet access.

In spite of infrastructural constraints, the University has been maintaining its academic excellence. The University has strictly adhered to the academic calendar, conducted the examinations and declared the results on time. The students from the University have found placements not only in State and Central Government Services, but also in various institutions, industries and organizations. Many students have emerged successful in the National Eligibility Test (NET).

Since inception, the University has made significant progress in teaching, research, innovations in curriculum development and developing infrastructure.

About IDE

The formal system of higher education in our country is facing the problems of access, limitation of seats, lack of facilities and infrastructure. Academicians from various disciplines opine that it is learning which is more important and not the channel of education. The education through distance mode is an alternative mode of imparting instruction to overcome the problems of access, infrastructure and socio-economic barriers. This will meet the demand for qualitative higher education of millions of people who cannot get admission in the regular system and wish to pursue their education. It also helps interested employed and unemployed men and women to continue with their higher education. Distance education is a distinct approach to impart education to learners who remained away in the space and/or time from the teachers and teaching institutions on account of economic, social and other considerations. Our main aim is to provide higher education opportunities to those who are unable to join regular academic and vocational education programmes in the affiliated colleges of the University and make higher education reach to the doorsteps in rural and geographically remote areas of Arunachal Pradesh in particular and North-eastern part of India in general. In 2008, the Centre for Distance Education has been renamed as "Institute of Distance Education (IDE)."

Continuing the endeavor to expand the learning opportunities for distant learners, IDE has introduced Post Graduate Courses in 5 subjects (Education, English, Hindi, History and Political Science) from the Academic Session 2013-14.

The Institute of Distance Education is housed in the Physical Sciences Faculty Building (first floor) next to the University Library. The University campus is 6 kms from NERIST point on National Highway 52A. The University buses ply to NERIST point regularly.

Outstanding Features of Institute of Distance Education:

(i) At Par with Regular Mode

Eligibility requirements, curricular content, mode of examination and the award of degrees are on par with the colleges affiliated to the Rajiv Gandhi University and the Department(s) of the University.

(ii) Self-Instructional Study Material (SISM)

The students are provided SISM prepared by the Institute and approved by Distance Education Council (DEC), New Delhi. This will be provided at the time of admission at the IDE or its Study Centres. SISM is provided only in English except Hindi subject.

(iii) Contact and Counselling Programme (CCP)

The course curriculum of every programme involves counselling in the form of personal contact programme of duration of approximately 7-15 days. The CCP shall not be compulsory for BA. However for professional courses and MA the attendance in CCP will be mandatory.

(iv) Field Training and Project

For professional course(s) there shall be provision of field training and project writing in the concerned subject.

(v) Medium of Instruction and Examination

The medium of instruction and examination will be English for all the subjects except for those subjects where the learners will need to write in the respective languages.

(vi) Subject/Counselling Coordinators

For developing study material, the IDE appoints subject coordinators from within and outside the University. In order to run the PCCP effectively Counselling Coordinators are engaged from the Departments of the University, The Counselling-Coordinators do necessary coordination for involving resource persons in contact and counselling programme and assignment evaluation. The learners can also contact them for clarifying their difficulties in then respective subjects.

SYLLABI-BOOK MAPPING TABLE

Mathematics and Statistics

Syllabi	Mapping in Book
UNIT I: Co-Ordinate Geometry (Two Dimensional) and Algebra Equation of a Straight Line: Slope, Intercept—Derivation of a Straight Line Given (a) Intercept and Slope and (b) Intercepts—Angle between Two Lines, Conditions of Line for Being Parallel. Circle: Derivation of the Equation of a Circle given a Point and Radius, Derivation of the Equation of a Parabola, Definition of Hyperbola and Ellipse, Binomial Expansion for a Positive, Negative or Fractional Exponent, Exponential and Logarithmic Series.	**Unit 1:** Coordinate Geometry and Algebra **(Pages 3-57)**
UNIT II: Matrix Algebra Scalar and Vector, Length of a Vector, Addition, Subtraction and Scalar Products of Two Vectors, Angle between Two Vectors, Cauchy-Schwarz Inequality, Vector Space and Normed Space, Basis of Vector Space, The Standard Basis, Spanning of Vector Space: Linear Combination and Linear Independence. Types of Matrices: Null, Unit and Idempotent Matrices, matrix Operations, Determinants, Matrix Inversion and Solution of Simultaneous Equations, Cramer's Rule, Rank of a Matrix, Characteristics Roots and Vectors.	**Unit 2:** Matrix Algebra **(Pages 59-135)**
UNIT III: Differentiation Limit and Continuity of Functions, Basic Rules of Differentiation, Partial and Total Differentiation, Indeterminate Form, L' Hopital Rules, Maxima and Minima, Points of Inflexion, Constrained Maximization and Minimization, Lagrangean Multiplier, Applications to Elasticity of Demand and Supply, Equilibrium to Consumer and Firm.	**Unit 3:** Differentiation **(Pages 137-201)**
UNIT IV: Integration Integral as Anti-Derivative, Basic Rules of Integration, Indefinite and Definite Integral, Beta and Gamma Functions, Improper Integral of $\int_0^\infty e^{-x^2}dx$, Application to Derivation of Total Revenue and Total Cost from Marginal Revenue and Marginal Cost, Estimation of Consumer Surplus and Producer Surplus, First Order Differential Equation.	**Unit 4:** Integration **(Pages 203-264)**
UNIT V: Linear Programming Concept, Objectives and Uses of Linear Programming in Economics, Graphical Method, Slack and Surplus Variables, Feasible Region and Basic Solution, Problem of Degeneration, Simplex Method, Solution of Primal and Dual Models.	**Unit 5:** Linear Programming **(Pages 265-311)**

UNIT VI: Probability
The Concept of Sample Space and Elementary Events, Mutually Exclusive Events, Dependent and Independent Events, Compound Events, A-Priori and Empirical Definition, Addition and Multiplication Theorems, Compound and Conditional Probability, Bayes' Theorem.

UNIT VII: Probability Distribution
Random Variable, Probability Function and Probability Density Function, Expectation, Variance, Covariance, Variance of a Linear Combination of Variables, Moments and Moment Generating Functions, Binomial, Poisson, Beta, Gamma and Normal Distributions, Derivation of Moments around Origin and Moments around Mean, Standard Normal Distribution.

UNIT VIII: Statistical Inference
Concept of Sampling Distribution, χ^2, t and F Distributions and their Properties, Type-I and Type-II Errors, One-Tailed and Two-Tailed Tests, Testing of Hypothesis based on Z, χ^2, t and F Distributions.

UNIT IX: Correlation and Regression
Simple Correlation and Its Properties, Range of Correlation Coefficient, Spearman's Rank Correlation (Tied and Untied).
Regression, OLS and its Assumptions, Estimation of Two Regression Lines, Angle Between Two Regression Lines, Properties of Regression Coefficients, Standard Error of Regression Coefficients, Partial and Multiple Coefficients, General Regression Model, Regression Coefficients and their Testing of Significance.

UNIT X: Index Numbers and Time Series
Index Number: Laspeyere's, Paasche's and Fisher's Index Numbers, Test for Ideal Index Numbers, Base Shifting, Base Splicing and Deflating, Concept of Constant Utility Index Number.
Time Series: Components of Time Series, Methods of Estimation of Linear and Non-Linear Trend.

CONTENTS

INTRODUCTION

NOTES

Mathematics is the study of quantity, structure, space and change. The mathematician, Benjamin Peirce called mathematics 'the science that draws necessary conclusions'. Hence, Mathematics is the most important subject for achieving excellence in any field of Science and Engineering. Mathematical statistics is the application of mathematics to statistics, which was originally conceived as the science of the state—the collection and analysis of facts about a country: its economy, land, military, population, and so forth. Mathematical techniques which are used for this include mathematical analysis, linear algebra, stochastic analysis, differential equations, and measure-theoretic probability theory.

Statistics is considered a mathematical science pertaining to the collection, analysis, interpretation or explanation and presentation of data. Statistical analysis is very important for taking decisions and is widely used by academic institutions, natural and social sciences departments, governments and business organizations. The word statistics is derived from the Latin word status which means a political state or government. It was originally applied in connection with kings and monarchs collecting data on their citizenry which pertained to state wealth, collection of taxes, study of population, and so on.

The subject of statistics is primarily concerned with making decisions about various disciplines of market and employment, such as stock market trends, unemployment rates in various sectors of industries, demographic shifts, interest rates, and inflation rates over the years, and so on. Statistics is also considered a science that deals with numbers or figures describing the state of affairs of various situations with which we are generally and specifically concerned. To a layman, it often refers to a column of figures or perhaps tables, graphs and charts relating to areas, such as population, national income, expenditures, production, consumption, supply, demand, sales, imports, exports, births, deaths, accidents, and so on.

This book, *Mathematics and Statistics*, has been designed keeping in mind the self-instruction mode format and follows a SIM pattern, wherein each unit begins with an 'Introduction' to the topic followed by the 'Unit Objectives'. The content is then presented in a simple and easy-to-understand manner, and is interspersed with 'Check Your Progress' questions to test the reader's understanding of the topic. 'Key Terms' and 'Summary' are useful tools for effective recapitulation of the text. A list of 'Questions and Exercises' is also provided at the end of each unit for effective recapitulation.

UNIT 6 PROBABILITY: BASIC CONCEPTS

NOTES

Structure

1.0 INTRODUCTION

In this unit, you will learn about the basic concepts of probability. Probability is the measure of the likeliness that an event will occur. The higher the probability of an event, the more certain we are that the event will occur. This unit will discuss the meaning of sample space, events, and so on. You will also learn about the addition and multiplication theorem on probability. In probability theory and statistics, Bayes' theorem (alternatively Bayes' law or Bayes' rule) describes the probability of an event, based on conditions that might be related to the event. The interpretation of Bayes' theorem depends on the interpretation of probability ascribed to the terms. You will learn about the Bayes theorem. Finally, you will learn about the random variable and probability distribution function. The three techniques for assigning probabilities to the values of the random variable, subjective probability assignment, a-priori probability assignment and empirical probability assignment, are also being discussed in this unit.

1.1 UNIT OBJECTIVES

After going through this unit, you will be able to:

- Discuss the basic rules of probability
- Explain dependent and independent events
- Discuss the significance of compound and conditional probability
- Understand Bayes' theorem
- Describe about random variable and probability distribution functions

1.2 PROBABILITY: BASICS

NOTES

The probability theory helps a decision-maker to analyse a situation and decide accordingly. The following are few examples of such situations:

- What is the *chance* that sales will increase if the price of the product is decreased?
- What is the *likelihood* that a new machine will increase productivity?
- How *likely* is it that a given project will be completed in time?
- What are the possibilities that a competitor will introduce a cheaper substitute in the market?

Probability theory is also called the theory of chance and can be mathematically derived using the standard formulas. A probability is expressed as a real number, $p \in [0, 1]$ and the probability number is expressed as a percentage (0 per cent to 100 per cent) and not as a decimal. For example, a probability of 0.55 is expressed as 55 per cent. When we say that the probability is 100 per cent, it means that the event is certain while the 0 per cent probability means that the event is impossible. We can also express probability of an outcome in the ratio format. For example, we have two probabilities, i.e., 'chance of winning' (1/4) and 'chance of not winning' (3/4), then using the mathematical formula of odds, we can say,

'Chance of winning' : 'Chance of not winning' = 1/4 : 3/4 = 1 : 3 or 1/3

We are using the probability in vague terms when we predict something for future. For example, we might say it will probably rain tomorrow or it will probably be a holiday the day after. This is subjective probability to the person predicting, but implies that the person believes the probability is greater than 50 per cent.

Different types of probability theories are as follows:

(i) Axiomatic Probability Theory

The axiomatic probability theory is the most general approach to probability, and is used for more difficult problems in probability. We start with a set of axioms, which serve to define a probability space. These axioms are not immediately intuitive and are developed using the classical probability theory.

(ii) Classical Theory of Probability

The classical theory of probability is the theory based on the number of favourable outcomes and the number of total outcomes. The probability is expressed as a ratio of these two numbers. The term 'favourable' is not the subjective value given to the outcomes, but is rather the classical terminology used to indicate that an outcome belongs to a given event of interest.

Classical Definition of Probability: If the number of outcomes belonging to an event E is N_E, and the total number of outcomes is N, then the probability of event E is defined as $p_E = \frac{N_E}{N}$.

For example, a standard pack of cards (without jokers) has 52 cards. If we randomly draw a card from the pack, we can imagine about each card as a possible outcome. Therefore, there are 52 total outcomes. Calculating all the outcome events and their probabilities, we have the following possibilities:

- Out of the 52 cards, there are 13 clubs. Therefore, if the event of interest is drawing a club, there are 13 favourable outcomes, and the probability of this event becomes $\frac{13}{52} = \frac{1}{4}$.
- There are 4 kings (one of each suit). The probability of drawing a king is $\frac{4}{52} = \frac{1}{13}$.
- What is the probability of drawing a king or a club? This example is slightly more complicated. We cannot simply add together the number of outcomes for each event separately (4 + 13 = 17) as this inadvertently counts one of the outcomes twice (the king of clubs). The correct answer is $\frac{16}{52}$ from $\frac{13}{52} + \frac{4}{52} - \frac{1}{52}$.

  We have this from the probability equation, P(club) + P(king) – P(king of clubs).
- Classical probability has limitations, because this definition of probability implicitly defines all outcomes to be equiprobable and this can be only used for conditions such as drawing cards, rolling dice, or pulling balls from urns. We cannot calculate the probability where the outcomes are unequal probabilities.

It is not that the classical theory of probability is not useful because of the described limitations. We can use this as an important guiding factor to calculate the probability of uncertain situations as just mentioned and to calculate the axiomatic approach to probability.

Frequency of Occurrence

This approach to probability is used for a wide range of scientific disciplines. It is based on the idea that the underlying probability of an event can be measured by repeated trials.

Probability as a Measure of Frequency: Let n_A be the number of times event A occurs after n trials. We define the probability of event A as,

$$P(A) = \lim_{n \to \infty} \frac{n_A}{n}$$

It is not possible to conduct an infinite number of trials. However, it usually suffices to conduct a large number of trials, where the standard of large depends on the probability being measured and how accurate a measurement we need.

Definition of Probability

To understand whether the sequence $\frac{n_A}{n}$ in the limit will converge to the same result every time, or it will not converge at all let us consider an experiment consisting of flipping a coin an infinite number of times. We want that the probability of heads must come up. The result may appear as the following sequence:

HTHHTTHHHHTTTTHHHHHHHHTTTTTTTTHHHHHHHHHHHHH HHHTTTTTTTTTTTTTTTT...

This shows that each run of k heads and k tails are being followed by another run of the same probability. For this example, the sequence $\frac{n_A}{n}$ oscillates between, $\frac{1}{3}$ and $\frac{2}{3}$ which does not converge. These sequences may be unlikely, and can be right. The definition given above does not express convergence in the required way, but it shows some kind of convergence in probability. The problem of exact formulation can be solved using the axiomatic probability theory.

NOTES

Empirical Probability Theory

The empirical approach to determine probabilities relies on data from actual experiments to determine approximate probabilities instead of the assumption of equal likeliness. Probabilities in these experiments are defined as the ratio of the frequency of the possibility of an event, $f(E)$, to the number of trials in the experiment, n, written symbolically as $P(E) = f(E)/n$. For example, while flipping a coin, the empirical probability of heads is the number of heads divided by the total number of flips.

The relationship between these empirical probabilities and the theoretical probabilities is suggested by the Law of Large Numbers. The law states that as the number of trials of an experiment increases, the empirical probability approaches the theoretical probability. Hence, if we roll a die a number of times, each number would come up approximately 1/6 of the time. The study of empirical probabilities is known as *statistics*.

1.2.1 Sample Space

A sample space is the collection of all possible events or outcomes of an experiment. For example, there are two possible outcomes of a toss of a fair coin: a head and a tail. Then, the sample space for this experiment denoted by S would be,

$$S = [H, T]$$

So that the probability of the sample space equals 1, or

$$P[S] = P[H,T] = 1$$

This is so because in the toss of the coin, either a head or a tail, must occur. Similarly, when we roll a die, any of the six faces can come as a result of the roll since there are a total of six faces. Hence, the sample space is $S = [1, 2, 3, 4, 5, 6]$, and $P[S] = 1$, since one of the six faces must occur.

1.2.2 Events

An event is an outcome or a set of outcomes of an activity or a result of a trial. For example, getting two heads in the trial of tossing three fair coins simultaneously would be an event. The following are the types of events:

- **Elementary Event:** An elementary event, also known as a simple event, is a single possible outcome of an experiment. For example, if we toss a fair coin, then the event of a head coming up is an elementary event. If the symbol for an elementary event is (E), then the probability of the event (E) is written as $P[E]$.
- **Joint Event:** A joint event, also known as a compound event, has two or more elementary events in it. For example, drawing a black ace from a pack of cards would be a joint event, since it contains two elementary events of black and ace.
- **Simple Probability:** Simple probability refers to a phenomenon where only a simple or elementary event occurs. For example, assume that event (E), the drawing of a diamond card from a pack of 52 cards, is a simple event. Since there are 13 diamond cards in the pack and each card is equally likely to be drawn, the probability of event (E) or $P[E] = 13/52$ or 1/4.

NOTES

- **Joint Probability:** The joint probability refers to the phenomenon of occurrence of two or more simple events. For example, assume that event (E) is a joint event (or compound event) of drawing a black ace from a pack of cards. There are two simple events involved in the compound event, which are, the card being black and the card being an ace. Hence, P[Black ace] or $P[E]$ = 2/52 since there are two black aces in the pack.
- **Complement of an Event:** The complement of any event A is the collection of outcomes that are not contained in A. This complement of A is denoted as A' (A prime). This means that the outcomes contained in A and the outcomes contained in A' must equal the total sample space. Therefore,

$$P[A] + P[A'] = P[S] = 1$$

or,

$$P[A] = 1 - P[A']$$

  For example, if a passenger airliner has 300 seats and it is nearly full, but not totally full, then event A would be the number of occupied seats and A' would be the number of unoccupied seats. Suppose there are 287 seats occupied by passengers and only 13 seats are empty. Typically, the stewardess will count the number of empty seats which are only 13 and report that 287 people are aboard. This is much simpler than counting 287 occupied seats. Accordingly, in such a situation, knowing event A' is much more efficient than knowing event A.
- **Mutually-Exclusive Events:** Two events are said to be mutually exclusive, if both events cannot occur at the same time as outcome of a single experiment. For example, if we toss a coin, then either event head or event tail would occur, but not both. Hence, these are mutually exclusive events.

Venn Diagrams

We can visualize the concept of events, their relationships and sample space using Venn diagrams. The sample space is represented by a rectangular region and the events and the relationships among these events are represented by circular regions within the rectangle.

For example, two mutually exclusive events A and B are represented in the Venn diagram in Figure 6.1.

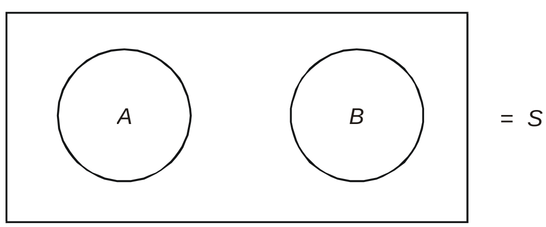


Fig. 6.1 *Venn Diagram of Two Mutually Exclusive Events A and B*

Event $P[A \cup B]$ is represented in the Venn diagram in Figure 6.2.

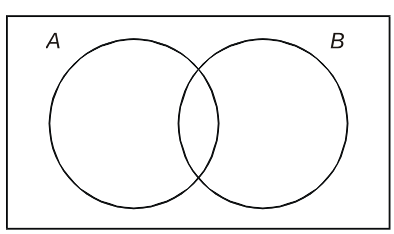


Fig. 6.2 *Venn Diagram Showing Event $P[A \cup B]$*

NOTES

Event [*AB*] is represented in Figure 6.3.

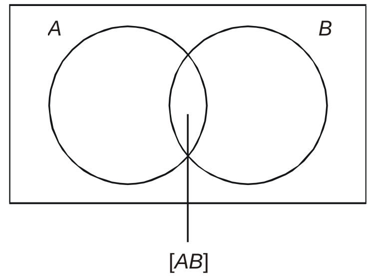


Fig. 6.3 *Venn Diagram Showing Event [A B]*

Union of Three Events

The process of combining two events to form the union can be extended to three events so that $P[A \cup B \cup C]$ would be the union of events *A*, *B*, and *C*. This union can be represented in a Venn diagram as in Figure 6.4. Example 6.1 explains the union of three events:

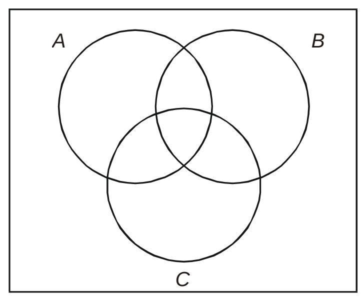


Fig. 6.4 *Venn Diagram Showing Union of Three Events* $P[A \cup B \cup C]$

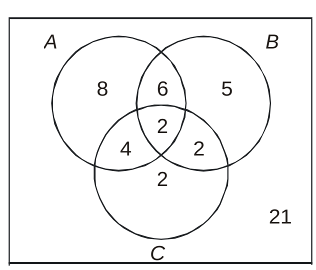
Example 6.1: A sample of 50 students is taken and a survey is made on the reading habits of the sample selected. The survey results are shown as follows:

Event	*Number of Students*	*Magazine They Read*
[*A*]	20	Time
[*B*]	15	Newsweek
[*C*]	10	Filmfare
[*AB*]	8	Time and Newsweek
[*AC*]	6	Time and Filmfare
[*BC*]	4	Newsweek and Filmfare
[*ABC*]	2	Time and Newsweek and Filmfare

Find out the probability that a student picked up at random from this sample of 50 students does not read any of these three magazines.

Solution:

The problem can be solved by a Venn diagram as follows:



Since there are 21 students who do not read any of the three magazines, the probability that a student picked up at random among this sample of 50 students who does not read any of these three magazines is 21/50.

The problem can also be solved by the formula for probability for union of three events, given as follows:

$$\begin{aligned} P[A \cup B \cup C] &= P[A] + P[B] + P[C] - P[AB] - P[AC] - P[BC] + P[ABC] \\ &= 20/50 + 15/50 + 10/50 - 8/50 - 6/50 - 4/50 + 2/50 \\ &= 29/50 \end{aligned}$$

The above is the probability that a student picked up at random among the sample of 50 reads either *Time* or *Newsweek* or *Filmfare* or any combination of the two or all the three. Hence, the probability that such a student does not read any of these three magazines is 21/50 which is [1 – 29/50].

1.2.3 Addition and Multiplication Theorem on Probability

This section will discuss the addition and multiplication theorem on probability.

Law of Addition

When two events are mutually exclusive, then the probability that either of the events will occur is the sum of their separate probabilities. For example, if you roll a single die, then the probability that it will come up with a face 5 or face 6, where event A refers to face 5 and event B refers to face 6 and both events being mutually exclusive events, is given by,

$$P[A \text{ or } B] = P[A] + P[B]$$

or,

$$\begin{aligned} P[5 \text{ or } 6] &= P[5] + P[6] \\ &= 1/6 + 1/6 \\ &= 2/6 = 1/3 \end{aligned}$$

$P[A \text{ or } B]$ is written as $P[A \cup B]$ and is known as P [A union B].

However, if events A and B are not mutually exclusive, then the probability of occurrence of either event A or event B or both is equal to the probability that event A occurs plus the probability that event B occurs minus the probability that events common to both A and B occur.

Symbolically, it can be written as,

$$P[A \cup B] = P[A] + P[B] - P[A \text{ and } B]$$

$P[A \text{ and } B]$ can also be written as $P[A \cap B]$, known as P [A intersection B] or simply $P[AB]$.

Events [A and B] consist of all those events which are contained in both A and B simultaneously. For example, in an experiment of taking cards out of a pack of 52 playing cards, assume the following:

Event A = An ace is drawn.

Event B = A spade is drawn.

Event [AB] = An ace of spade is drawn.

Hence,

$$\begin{aligned} P[A \cup B] &= P[A] + P[B] - P[AB] \\ &= 4/52 + 13/52 - 1/52 \\ &= 16/52 \\ &= 4/13 \end{aligned}$$

This is so, because there are 4 aces, 13 cards of spades, including 1 ace of spades out of a total of 52 cards in the pack. The logic behind subtracting $P[AB]$ is that the ace of spades is counted twice—once in event A (4 aces) and once again in event B (13 cards of spade including the ace).

NOTES

Another example for $P[A \cup B]$, where event A and event B are not mutually exclusive is as follows:

Suppose a survey of 100 persons revealed that 50 persons read *India Today* and 30 persons read *Time* magazine and 10 of these 100 persons read both *India Today* and *Time*. Then,

$$\text{Event } [A] = 50$$

$$\text{Event } [B] = 30$$

$$\text{Event } [AB] = 10$$

Since event $[AB]$ of 10 is included twice, both in event A as well as in event B, event $[AB]$ must be subtracted once in order to determine the event $[A \cup B]$ which means that a person reads *India Today* or *Time* or both. Hence,

$$P[A \cup B] = P[A] + P[B] - P[AB]$$

$$= 50/100 + 30/100 - 10/100$$

$$= 70/100 = 0.7$$

Law of Multiplication

Multiplication rule is applied when it is necessary to compute the probability if both events A and B will occur at the same time. The multiplication rule is different if the two events are independent as against the two events being not independent.

If events A and B are independent events, then the probability that they both will occur is the product of their separate probabilities. This is a strict condition so that events A and B are independent if and only if,

$$P[AB] = P[A] \times P[B]$$

or $$= P[A]\,P[B]$$

For example, if we toss a coin twice, then the probability that the first toss results in a head and the second toss results in a tail is given by,

$$P[HT] = P[H] \times P[T]$$

$$= 1/2 \times 1/2 = 1/4$$

However, if events A and B are not independent, meaning that the probability of occurrence of an event is dependent or conditional upon the occurrence or non-occurrence of the other event, then the probability that they will both occur is given by,

$$P[AB] = P[A] \times P[B/\text{Given outcome of } A]$$

This relationship is written as,

$$P[AB] = P[A] \times P[B/A] = P[A]\,P[B/A]$$

Where, $P[B/A]$ means the probability of event B on the condition that event A has occurred. As an example, assume that a bowl has 6 black balls and 4 white balls. A ball is drawn at random from the bowl. Then a second ball is drawn without replacement of the first ball back in the bowl. The probability of the second ball being black or white would depend upon the result of the first draw as to whether the first ball was black or white. The probability that both these balls are black is given by,

P [Two black balls] = *P* [Black on 1st draw] × *P* [Black on 2nd draw/Black on 1st draw]

= 6/10 × 5/9 = 30/90 = 1/3

This is so, because there are 6 black balls out of a total of 10, but if the first ball drawn is black then we are left with 5 black balls out of a total of 9 balls.

1.2.4 Independent Events

Two events *A* and *B* are said to be independent events, if the occurrence of one event is not influenced at all by the occurrence of the other. For example, if two fair coins are tossed, then the result of one toss is totally independent of the result of the other toss. The probability that a head will be the outcome of any one toss will always be 1/2, irrespective of whatever the outcome is of the other toss. Hence, these two events are independent.

Let us assume that one fair coin is tossed 10 times and it happens that the first nine tosses resulted in heads. What is the probability that the outcome of the tenth toss will also be a head? There is always a psychological tendency to think that a tail would be more likely in the tenth toss since the first nine tosses resulted in heads. However, since the events of tossing a coin 10 times are all independent events, the earlier outcomes have no influence whatsoever on the result of the tenth toss. Hence, the probability that the outcome will be a head on the tenth toss is still 1/2.

On the other hand, consider drawing two cards from a pack of 52 playing cards. The probability that the second card will be an ace would depend upon whether the first card was an ace or not. Hence, these two events are not independent events.

1.2.5 Conditional Probability

In many situations, a manager may know the outcome of an event that has already occurred and may want to know the chances of a second event occurring based upon the knowledge of the outcome of the earlier event. We are interested in finding out as to how additional information obtained as a result of the knowledge about the outcome of an event affects the probability of the occurrence of the second event. For example, let us assume that a new brand of toothpaste is being introduced in the market. Based on the study of competitive markets, the manufacturer has some idea about the chances of its success. Now, he introduces the product in a few selected stores in a few selected areas before marketing it nationally. A highly positive response from the test-market area will improve his confidence about the success of his brand nationally. Accordingly, the manufacturer's assessment of high probability of sales for his brand would be conditional upon the positive response from the test-market.

Let there be two events *A* and *B*. Then the probability that event *A* occurs, given that event *B* has occured. The notation is given by,

$$P[A/B] = \frac{P[AB]}{P[B]}$$

Where *P*[*A*/*B*] is interpreted as the probability of event *A* on the condition that event *B* has occurred and *P* [*AB*] is the joint probability of event *A* and event *B*, and *P*[*B*] is not equal to zero.

NOTES

As an example, let us suppose that we roll a die and we know that the number that came up is larger than 4. We want to find out the probability that the outcome is an even number given that it is larger than 4.

Let, Event A = Even

and Event B = Larger than 4

Then, $$\text{P[Even / Larger than 4]} = \frac{P[\text{Even and larger than 4}]}{P[\text{Larger than 4}]}$$

or $$P[A/B] = \frac{P[AB]}{P[B]} = (1/6)/(2/6) = 1/2$$

For, however independent events, $P[AB] = P[A]\,P[B]$. Thus, substituting this relationship in the formula for conditional probability, we get,

$$P[A/B] = \frac{P[AB]}{P[B]} = \frac{P[A]P[B]}{P[B]} = P[A]$$

This means that $P[A]$ will remain the same no matter what the outcome of event B is. For example, if we want to find out the probability of a head on the second toss of a fair coin, given that the outcome of the first toss was a head, this probability would still be 1/2 because the two events are independent events and the outcome of the first toss does not affect the outcome of the second toss.

1.3 BAYES' THEOREM

Reverend Thomas Bayes (1702–1761), introduced his theorem on probability, which is concerned with a method for estimating the probability of causes which are responsible for the outcome of an observed effect. Being a religious preacher himself as well as a mathematician, his motivation for the theorem came from his desire to prove the existence of God by looking at the evidence of the world that God created. He was interested in drawing conclusions about the causes by observing the consequences. The theorem contributes to the statistical decision theory in revising prior probabilities of outcomes of events based upon the observation and analysis of additional information.

Bayes' theorem makes use of conditional probability formula where the condition can be described in terms of the additional information which would result in the revised probability of the outcome of an event.

Suppose that, there are 50 students in our statistics class out of which 20 are male students and 30 are female students. Out of the 30 females, 20 are Indian students and 10 are foreign students. Out of the 20 male students, 15 are Indians and 5 are foreigners, so that out of all the 50 students, 35 are Indians and 15 are foreigners. This data can be presented in a tabular form as follows:

	Indian	***Foreigner***	***Total***
Male	15	5	20
Female	20	10	30
Total	35	15	50

Check Your Progress

1. List the different types of probability theories.
2. On what the classical theory of probability is based?
3. What is the Law of Large Numbers (LLN)?

NOTES

Based upon this information, the probability that a student picked up at random will be female is 30/50 or 0.6, since there are 30 females in the total class of 50 students. Now suppose that we are given additional information that the person picked up at random is Indian, then what is the probability that this person is a female? This additional information will result in revised probability or posterior probability in the sense that it is assigned to the outcome of the event after this additional information is made available.

Since we are interested in the revised probability of picking a female student at random provided that we know that the student is Indian. Let A_1 be the event female, A_2 be the event male and B be the event Indian. Then based upon our knowledge of conditional probability, the Bayes' theorem can be stated as,

$$P(A_1 / B) = \frac{P(A_1)P(B / A_1)}{P(A_1)P(B / A_1) + P(A_2)(P(B / A_2)}$$

In the example discussed, there are two basic events which are A_1 (female) and A_2 (male). However, if there are *n* basic events, $A_1, A_2, ...A_n$, then Bayes' theorem can be generalized as,

$$P(A_1 / B) = \frac{P(A_1)P(B / A_1)}{P(A_1)P(B / A_1) + P(A_2)(P(B / A_2) + ... + P(A_n)P(B / A_n)}$$

Solving the case of two events we have,

$$P(A_1 / B) = \frac{(30 / 50)(20 / 30)}{(30 / 50)(20 / 30) + (20 / 50)(15 / 20)} = 20 / 35 = 4 / 7 = 0.57$$

This example shows that while the prior probability of picking up a female student is 0.6, the posterior probability becomes 0.57 after the additional information that the student is an American is incorporated in the problem.

Refer Example 6.2 to understand the theorem better.

Example 6.2: A businessman wants to construct a hotel in New Delhi. He generally builds three types of hotels. These are hotels with 50 rooms, 100 rooms and 150 rooms, depending upon the demand for rooms, which is a function of the area in which the hotel is located, and the traffic flow. The demand can be categorized as low, medium or high. Depending upon these various demands, the businessman has made some preliminary assessment of his net profits and possible losses (in thousands of dollars) for these various types of hotels. These pay-offs are shown in the following table:

		Demand for Rooms			
		Low (A_1)	*Medium* (A_2)	*High* (A_3)	
		0.2	0.5	0.3	Demand Probability
Number of Rooms	R_1=(50)	25	35	50	
	R_2=(100)	–10	40	70	
	R_3=(150)	–30	20	100	

Solution: The businessman has also assigned 'prior probabilities' to the demand structure or rooms. These probabilities reflect the initial judgement of the businessman based upon his intuition and his degree of belief regarding the outcomes of the states of nature.

Demand for Rooms	*Probability of Demand*
Low (A_1)	0.2
Medium (A_2)	0.5
High (A_3)	0.3

NOTES

Based upon these values, the expected pay-offs for various rooms can be computed as,

$$EV(50) = (25 \times 0.2) + (35 \times 0.5) + (50 \times 0.3) = 37.50$$

$$EV(100) = (-10 \times 0.2) + (40 \times 0.5) + (70 \times 0.3) = 39.00$$

$$EV(150) = (-30 \times 0.2) + (20 \times 0.5) + (100 \times 0.3) = 34.00$$

This gives us the maximum pay-off of $39,000 for building a 100 rooms hotel.

Now, the hotelier must decide whether to gather additional information regarding the states of nature, so that these states can be predicted more accurately than the preliminary assessment. The basis of such a decision would be the cost of obtaining additional information. If this cost is less than the increase in maximum expected profit, then such additional information is justified.

Suppose that the businessman asks a consultant to study the market and predict the states of nature more accurately. This study is going to cost the businessman $10,000. This cost would be justified if the maximum expected profit with the new states of nature is at least $10,000 more than the expected pay-off with the prior probabilities. The consultant made some studies and came up with the estimates of low demand (X_1), medium demand (X_2), and high demand (X_3) with a degree of reliability in these estimates. This degree of reliability is expressed as conditional probability which is the probability that the consultant's estimate of low demand will be correct and the demand will be actually low. Similarly, there will be a conditional probability of the consultant's estimate of medium demand, when the demand is actually low and, so on. These conditional probabilities are expressed in the Table 6.1.

Table 6.1 *Conditional Probabilities*

		X_1	X_2	X_3
States of	(A_1)	0.5	0.3	0.2
Nature	(A_2)	0.2	0.6	0.2
(Demand)	(A_3)	0.1	0.3	0.6

The values in the preceding table are conditional probabilities and are interpreted as follows:

The first value of 0.5 is the probability that the consultant's prediction will be for low demand (X_1) when the demand is actually low. Similarly, the probability is 0.3 that the consultant's estimate will be for medium demand (X_2) when in fact the demand is low, and so on. In other words, $P(X_1/A_1) = 0.5$ and $P(X_2/A_1) = 0.3$. Similarly, $P(X_1/A_2) = 0.2$ and $P(X_2/A_2) = 0.6$, and so on.

Our objective is to obtain posteriors which are computed by taking the additional information into consideration. One way to reach this objective is to first compute the joint probability, which is the product of prior probability and conditional probability for each state of nature. Joint probabilities as computed is given as, Joint Probabilities

State of Nature	Prior Probability	$P(A_1X_1)$	Joint Probabilities $P(A_1X_2)$	$P(A_1X_3)$
A_1	0.2	$0.2 \times 0.5 = 0.10$	$0.2 \times 0.3 = 0.06$	$0.2 \times 0.2 = 0.04$
A_2	0.5	$0.5 \times 0.2 = 0.10$	$0.5 \times 0.6 = 0.30$	$0.5 \times 0.2 = 0.10$
A_3	0.3	$0.3 \times 0.1 = 0.03$	$0.3 \times 0.3 = 0.09$	$0.3 \times 0.6 = 0.18$
Total Marginal Probabilities.		$= 0.23$	$= 0.45$	$= 0.32$

NOTES

Now, the posterior probabilities for each state of nature A_i are calculated as,

$$P(A_i / X_j) = \frac{\text{Joint probability of } A_i \text{ and } X_j}{\text{Marginal probability of } X_i}$$

By using this formula, the joint probabilities are converted into posterior probabilities and the computed table for these posterior probabilities is given as,

States of Nature	Posterior Probabilities $P(A_1/X_1)$	$P(A_1/X_2)$	$P(A_1/X_3)$
A_1	$0.1/0.23 = 0.435$	$0.06/0.45 = 0.133$	$0.04/0.32 = 0.125$
A_2	$0.1/0.23 = 0.435$	$0.30/0.45 = 0.667$	$0.1/0.32 = 0.312$
A_3	$0.03/0.23 = 0.130$	$0.09/0.45 = 0.200$	$0.18/0.32 = 0.563$
Total	$= 1.0$	$= 1.0$	$= 1.0$

Now, we have to compute the expected pay-offs for each course of action with the new posterior probabilities assigned to each state of nature. The net profits for each course of action for a given state of nature is the same as before and is restated. These net profits are expressed in thousands of dollars.

		Low (A_1)	Medium (A_2)	High (A_3)
Number of Rooms	(R_1)	25	35	50
	(R_2)	– 10	40	70
	(R_3)	– 30	20	100

Let O_{ij} be the monetary outcome of course of action i when j is the corresponding state of nature, so that in the above case O_{i1} will be the outcome of course of action R_1 and state of nature A_1, which in our case is \$25,000. Similarly, O_{i2} will be the outcome of action R_2 and state of nature A_2, which in our case is \$10,000, and so on. The expected value EV (in thousands of dollars) is calculated on the basis of the actual state of nature that prevails as well as the estimate of the state of nature as provided by the consultant. These expected values are calculated as,

Course of action $= R_i$
Estimate of consultant $= X_i$
Actual state of nature $= A_i$

Where, $i = 1, 2, 3$

Then,

(i) Course of action $= R_1$ = Build 50 rooms hotel

$$EV\left(\frac{R_1}{X_1}\right) = \Sigma P\left(\frac{A_i}{X_1}\right) O_{i1}$$

$$= 0.435(25) + 0.435(-10) + 0.130(-30)$$

$$= 10.875 - 4.35 - 3.9 = 2.625$$

$$EV\left(\frac{R_1}{X_2}\right) = \Sigma P\left(\frac{A_i}{X_2}\right)O_{i1}$$

$$= 0.133(25) + 0.667(-10) + 0.200(-30)$$

$$= 3.325 - 6.67 - 6.0 = -9.345$$

$$EV\left(\frac{R_1}{X_3}\right) = \Sigma P\left(\frac{A_i}{X_3}\right)O_{i1}$$

$$= 0.125(25) + 0.312(-10) + 0.563(-30)$$

$$= 3.125 - 3.12 - 16.89$$

$$= -16.885$$

(ii) Course of action = R_2 = Build 100 rooms hotel

$$EV\left(\frac{R_2}{X_1}\right) = \Sigma P\left(\frac{A_i}{X_1}\right)O_{i2}$$

$$= 0.435(35) + 0.435(40) + 0.130(20)$$

$$= 15.225 + 17.4 + 2.6 = 35.225$$

$$EV\left(\frac{R_2}{X_2}\right) = \Sigma P\left(\frac{A_i}{X_1}\right)O_{i2}$$

$$= 0.133(35) + 0.667(40) + 0.200(20)$$

$$= 4.655 + 26.68 + 4.0 = 35.335$$

$$EV\left(\frac{R_2}{X_3}\right) = \Sigma P\left(\frac{A_i}{X_3}\right)O_{i2}$$

$$= 0.125(35) + 0.312(40) + 0.563(20)$$

$$= 4.375 + 12.48 + 11.26 = 28.115$$

(iii) Course of action = R_3 = Build 150 rooms hotel

$$EV\left(\frac{R_3}{X_1}\right) = \Sigma P\left(\frac{A_i}{X_1}\right)O_{i3}$$

$$= 0.435(50) + 0.435(70) + 0.130(100)$$

$$= 21.75 + 30.45 + 13 = 65.2$$

$$EV\left(\frac{R_3}{X_2}\right) = \Sigma P\left(\frac{A_i}{X_2}\right)O_{i3}$$

$$= 0.133(50) + 0.667(70) + 0.200(100)$$

$$= 6.65 + 46.69 + 20 = 73.34$$

$$EV\left(\frac{R_3}{X_3}\right) = \Sigma P\left(\frac{A_i}{X_3}\right)O_{i3}$$

$$= 0.125(50) + 0.312(70) + 0.563(100)$$
$$= 6.25 + 21.84 + 56.3 = 84.39$$

The expected values in thousands of dollars, as calculated, are presented as follows in a tabular form.

NOTES

Expected Posterior Pay-Offs

Outcome	$EV(R_1/X_i)$	$EV(R_2/X_i)$	$EV(R_3/X_i)$
X_1	2.625	35.225	65.2
X_2	–9.345	35.335	73.34
X_3	–16.885	28.115	84.39

This table can now be analysed. If the outcome is X_1, it is desirable to build 150 rooms hotel, since the expected pay-off for this course of action is maximum of $65,200. Similarly, if the outcome is X_2, the course of action should again be R_3 since the maximum pay-off is $73,34. Finally, if the outcome is X_3, the maximum payoff is $84,390 for course of action R_3.

Accordingly, given these conditions and the pay-off, it would be advisable to build a 150 rooms hotel.

1.4 RANDOM VARIABLE AND PROBABILITY DISTRIBUTION FUNCTIONS

This section will discuss about random variable and probability distribution functions.

1.4.1 Random Variable

A random variable takes on different values as a result of the outcomes of a random experiment. In other words, a function which assigns numerical values to each element of the set of events that may occur (i.e., every element in the sample space) is termed as a random variable. The value of a random variable is the general outcome of the random experiment. One should always make a distinction between the random variable and the values that it can take on. All these can be illustrated by a few examples as shown in Table 6.2.

Table 6.2 *Random Variable*

Random Variable	*Values of the Random Variable*	*Description of the Values of the Random Variable*
X	0, 1, 2, 3, 4	Possible number of heads in four tosses of a fair coin
Y	1, 2, 3, 4, 5, 6	Possible outcomes in a single throw of a die
Z	2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12	Possible outcomes from throwing a pair of dice
M	0, 1, 2, 3, ... S	Possible sales of newspapers by a newspaper boy, S representing his stock

NOTES

All the stated random variable assignments cover every possible outcome and each numerical value represents a unique set of outcomes. A random variable can be either discrete or continuous. If a random variable is allowed to take on only a limited number of values, it is a discrete random variable, but if it is allowed to assume any value within a given range, it is a continuous random variable. Random variables presented in Table 6.2 are examples of discrete random variables. We can have continuous random variables if they can take on any value within a range of values, for example, within 2 and 5, in that case we write the values of a random variable x as,

$$2 \leq x \leq 5$$

Techniques of Assigning Probabilities

We can assign probability values to the random variables. Since the assignment of probabilities is not an easy task, we should observe certain following rules in this context:

- A probability cannot be less than zero or greater than one, i.e., $0 \leq pr \leq 1$, where pr represents probability.
- The sum of all the probabilities assigned to each value of the random variable must be exactly one.

There are three techniques of assignment of probabilities to the values of the random variable that are as follows:

(i) ***Subjective Probability Assignment:*** It is the technique of assigning probabilities on the basis of personal judgement. Such assignment may differ from individual to individual and depends upon the expertise of the person assigning the probabilities. It cannot be termed as a rational way of assigning probabilities, but is used when the objective methods cannot be used for one reason or the other.

(ii) ***A-Priori Probability Assignment:*** It is the technique under which the probability is assigned by calculating the ratio of the number of ways in which a given outcome can occur to the total number of possible outcomes. The basic underlying assumption in using this procedure is that every possible outcome is likely to occur equally. However, at times the use of this technique gives ridiculous conclusions. For example, we have to assign probability to the event that a person of age 35 will live upto age 36. There are two possible outcomes, he lives or he dies. If the probability assigned in accordance with a-priori probability assignment is half then the same may not represent reality. In such a situation, probability can be assigned by some other techniques.

(iii) ***Empirical Probability Assignment:*** It is an objective method of assigning probabilities and is used by the decision-makers. Using this technique the probability is assigned by calculating the relative frequency of occurrence of a given event over an infinite number of occurrences. However, in practice only a finite (perhaps very large) number of cases are observed and relative frequency of the event is calculated. The probability assignment through this technique may as well be unrealistic, if future conditions do not happen to be a reflection of the past.

Thus, what constitutes the 'best' method of probability assignment can only be judged in the light of what seems best to depict reality. It depends upon the nature of the problem and also on the circumstances under which the problem is being studied.

1.4.2 Probability Distribution Functions: Discrete and Continuous

NOTES

When a random variable x takes discrete values $x_1, x_2,...., x_n$ with probabilities $p_1, p_2,...,p_n$, we have a discrete probability distribution of X.

The function $p(x)$ for which $X = x_1, x_2,..., x_n$ takes values $p_1, p_2,... ,p_n$, is the probability function of X.

The variable is discrete because it does not assume all values. Its properties are:

$$p(x_i) = \text{Probability that } X \text{ assumes the value } x$$
$$= \text{Prob } (x = x_i) = p_i$$
$$p(x) \geq 0, \Sigma p(x) = 1$$

For example, four coins are tossed and the number of heads X noted. X can take values 0, 1, 2, 3, 4 heads.

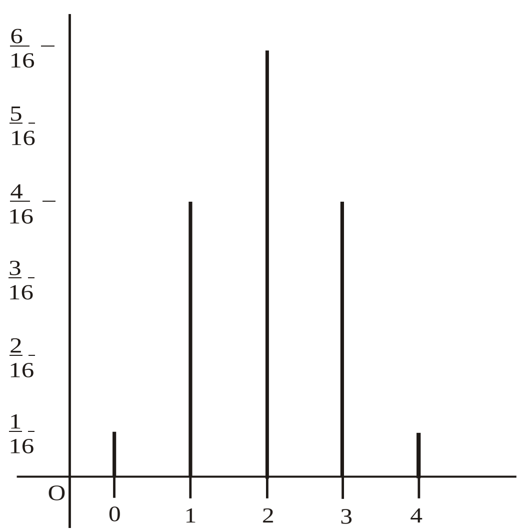
$$p(X = 0) = \left(\frac{1}{2}\right)^4 = \frac{1}{16}$$

$$p(X = 1) = {}^4C_1 \left(\frac{1}{2}\right)\left(\frac{1}{2}\right)^3 = \frac{4}{16}$$

$$p(X = 2) = {}^4C_2 \left(\frac{1}{2}\right)^2 \left(\frac{1}{2}\right)^2 = \frac{6}{16}$$

$$p(X = 3) = {}^4C_3 \left(\frac{1}{2}\right)^3 \left(\frac{1}{2}\right) = \frac{4}{16}$$

$$p(X = 4) = {}^4C_4 \left(\frac{1}{2}\right)^4 \left(\frac{1}{2}\right)^0 = \frac{1}{16}$$



$$\sum_{x=2}^{4} p(x) = \frac{1}{16} + \frac{4}{16} + \frac{6}{16} + \frac{4}{16} + \frac{1}{16} = 1$$

This is a discrete probability distribution (refer Example 6.3).

NOTES

Example 6.3: If a discrete variable X has the following probability function, then find (*i*) a (*ii*) $p(X \leq 3)$ (*iii*) $p(X \geq 3)$.

Solution:

x_1	$p(x_i)$
0	0
1	a
2	$2a$
3	$2a^2$
4	$4a^2$
5	$2a$

Since $\Sigma p(x) = 1$, $0 + a + 2a + 2a^2 + 4a^2 + 2a = 1$

$\therefore \quad 6a^2 + 5a - 1 = 0$, so that $(6a - 1)(a + 1) = 0$

$$a = \frac{1}{6} \text{ or } a = -1 \text{ (not admissible)}$$

For $a = \frac{1}{6}$, $p(X \leq 3) = 0 + a + 2a + 2a^2 = 2a^2 + 3a = \frac{5}{9}$

$$p(X \geq 3) = 4a^2 + 2a = \frac{4}{9}$$

Discrete Distributions

There are several discrete distributions. Some other discrete distributions are described as follows:

(i) Uniform or Rectangular Distribution

Each possible value of the random variable x has the same probability in the uniform distribution. If x takes vaues $x_1, x_2 ......, x_k$, then,

$$p(x_i, k) = \frac{1}{k}$$

The numbers on a die follow the uniform distribution,

$$p(x_i, 6) = \frac{1}{6} \text{ (Here, } x = 1, 2, 3, 4, 5, 6)$$

Bernoulli Trials

In a Bernoulli experiment, an even E either happens or does not happen (E'). Examples are, getting a head on tossing a coin, getting a six on rolling a die, and so on.

The Bernoulli random variable is written,

$$X = 1 \text{ if } E \text{ occurs}$$
$$= 0 \text{ if } E' \text{ occurs}$$

Since there are two possible values it is a case of a discrete variable where,

Probability of success $= p = p(E)$

Profitability of failure = $1 - p = q = \text{p}(E')$

We can write,

For $k = 1$, $f(k) = p$

For $k = 0$, $f(k) = q$

For $k = 0$ or 1, $f(k) = p^k q^{1-k}$

NOTES

Negative Binomial

In this distribution, the variance is larger than the mean.

Suppose, the probability of success p in a series of independent Bernoulli trials remains constant.

Suppose the rth success occurs after x failures in $x + r$ trials, then

(i) The probability of the success of the last trial is p.

(ii) The number of remaining trials is $x + r - 1$ in which there should be $r - 1$ successes. The probability of $r - 1$ successes is given by,

$${}^{x+r-1}C_{r-1}\, p^{r-1}\, q^{x}$$

The combined pobability of Cases (i) and (ii) happening together is,

$$p(x) = px^{x+r-1}C_{r-1}\, p^{r-1}\, q^{x} \qquad x = 0, 1, 2, ...$$

This is the Negative Binomial distribution. We can write it in an alternative form,

$$p(x) = {}^{-r}C_x\, p^r\, (-q)^x \qquad x = 0, 1, 2, ...$$

This can be summed up as follows:

In an infinite series of Bernoulli trials, the probability that $x + r$ trials will be required to get r successes is the negative binomial,

$$p(x) = {}^{x+r-1}C_{r-1}\, p^{r-1}\, q^{x} \qquad r \geq 0$$

If $r = 1$, it becomes the geometric distribution.

If $p \to 0$, $\to \infty$, $rp = m$ a constant, then the negative binomial tends to be the Poisson distribution.

(ii) Geometric Distribution

Suppose the probability of success p in a series of independent trials remains constant.

Suppose, the first success occurs after x failures, i.e., there are x failures preceding the first success. The probability of this event will be given by $p(x) = q^x p$ ($x = 0, 1, 2, ...$)

This is the geometric distribution and can be derived from the negative binomial. If we put $r = 1$ in the negative binomial distribution, then

$$p(x) = {}^{x+r-1}C_{r-1}\, p^{r-1}\, q^{x}$$

We get the geometric distribution,

$$p(x) = {}^{x}C_0\, p^1\, q^x = pq^x$$

$$\Sigma p(x) = \sum_{n=0}^{p} q^{x} p = \frac{p}{1-q} = 1$$

NOTES

$$E(x) = \text{Mean} = \frac{p}{q}$$

$$\text{Variance} = \frac{p}{q^2}$$

$$\text{Mode} = \left(\frac{1}{2}\right)^x$$

Refer Example 6.4 to understand it better.

Example 6.4: Find the expectation of the number of failures preceding the first success in an infinite series of independent trials with constant probability p of success.

Solution:

The probability of success in,

$$\text{1st trial} = p \text{ (Success at once)}$$
$$\text{2nd trial} = qp \text{ (One failure, then success, and so on)}$$
$$\text{3rd trial} = q^2p \text{ (Two failures, then success, and so on)}$$

The expected number of failures preceding the success,

$$E(x) = 0 \,.\, p + 1.\, pq + 2p^2p + \text{............}$$
$$= pq(1 + 2q + 3q^2 + \text{.........})$$
$$= pq\frac{1}{(1-q)^2} = qp\frac{1}{p^2} = \frac{q}{p}$$

Since $p = 1 - q$.

(iii) Hypergeometic Distribution

From a finite population of size N, a sample of size n is drawn without replacement.

Let there be N_1 successes out of N.

The number of failures is $N_2 = N - N_1$.

The disribution of the random variable X, which is the number of successes obtained in the discussed case, is called the hypergeometic distribution.

$$p(x) = \frac{{}^{N_1}C_x\,{}^{N}C_{n-x}}{{}^{N}C_n} \quad (X = 0, 1, 2, ..., n)$$

Here, x is the number of successes in the sample and $n - x$ is the number of failures in the sample.

It can be shown that,

$$\text{Mean} : E(X) = n\frac{N_1}{N}$$

$$\text{Variance} : \text{Var (X)} = \frac{N-n}{N-1}\left(\frac{nN_1}{N} \quad \frac{nN_1^2}{N}\right)$$

Example 6.5: There are 20 lottery tickets with three prizes. Find the probability that out of 5 tickets purchased exactly two prizes are won.

Solution:

We have $N_1 = 3$, $N_2 = N - N_1 = 17$, $x = 2$, $n = 5$.

$$p(2) = \frac{{}^3C_2 \; {}^{17}C_3}{{}^{20}C_5}$$

The probability of no prize $p(0) = \frac{{}^3C_0 \; {}^{17}C_5}{{}^{20}C_5}$

The probability of exactly 1 prize $p(1) = \frac{{}^3C_1 \; {}^{17}C_4}{{}^{20}C_5}$

Example 6.6: Examine the nature of the distibution if balls are drawn, one at a time without replacement, from a bag containing m white and n black balls.

Solution:

It is the hypergeometric distribution. It corresponds to the probability that x balls will be white out of r balls so drawn and is given by,

$$p(x) = \frac{{}^xC_x \; {}^nC_{r-x}}{{}^{m+n}C_r}$$

(iv) Multinomial

There are k possible outcomes of trials, viz., $x_1, x_2, ..., x_k$ with probabilities $p_1, p_2, ..., p_k$, n independent trials are performed. The multinomial distibution gives the probability that out of these n trials, x_1 occurs n_1 times, x_2 occurs n_2 times, and so on. This is given by

$$\frac{n!}{n_1!n_2!....n_k!} p_1^{n_1} p_2^{n_2} \cdots p_k^{n}$$

Where, $\sum_{i=1}^{k} n_i = n$

Characteristic Features of the Binomial Distribution

The following are the characteristic features of binomial distribution:

(i) It is a discrete distribution.

(ii) It gives the probability of x successes and $n - x$ failures in a specific order.

(iii) The experiment consists of n repeated trials.

(iv) Each trial results in a success or a failure.

(v) The probability of success remains constant from trial to trial.

(vi) The trials are independent.

(vii) The success probability p of any outcome remains constant over time. This condition is usually not fully satisfied in situations involving management and economics, e.g., the probability of response from successive informants is not the same. However, it may be assumed that the condition is reasonably well satisfied in many cases and that the outcome of one trial does not depend on the outcome of another. This condition too, may not be fully satisfied in many cases. An investigator may not approach a second informant with the same mind set as used for the first informant.

NOTES

(viii) The binomial distribution depends on two parameters, n and p. Each set of different values of n, p has a different binomial distribution.

(ix) If $p = 0.5$, the distribution is *symmetrical.* For a symmetrical distribution, in n

$$\text{Prob } (X = 0) = \text{Prob } (X = n)$$

i.e., the probabilities of 0 or n successes in n trials will be the same. Similarly,

$$\text{Prob } (X = 1) = \text{Prob}(X = n - 1), \text{ and so on.}$$

If $p > 0.5$, the distribution is not symmetrical. The probabilities on the right are larger than those on the left. The reverse case is when $p < 0.5$.

When n becomes large the distribution becomes bell shaped. Even when n is not very large but $p \cong 0.5$, it is fairly bell shaped.

(x) The binomial distribution can be approximated by the normal. As n becomes large and p is close to 0.5, the approximation becomes better.

Through the following examples you can understand multinomial better.

Example 6.7: Explain the concept of a discrete probability distribution.

Solution: If a random variable x assumes n discrete values $x_1, x_2, ...x_n$, with respective probabilities $p_1, p_2, ...p_n (p_1 + p_2 + ... + p_n = 1)$, then the distribution of values x_i with probabilities p_i (= 1, 2,.....n), is called the discrete probability distribution of x.

The frequency function or frequency distribution of x is defined by $p(x)$ which for different values $x_1, x_2, ...x_n$ of x, gives the corresponding probabilities:

$$p(x_i) = p_i \text{ where, } p(x) \geq 0 \ \Sigma p(x) = 1$$

Example 6.8: For the following probability distribution, find $p(x > 4)$ and $p(x \geq 4)$:

x	0	1	2	3	4	5
$p(x)$	0	a	$a/2$	$a/2$	$a/4$	$a/4$

Solution:

Since, $\Sigma p(x) = 1, 0 + a + \frac{a}{2} + \frac{a}{2} + \frac{a}{4} + \frac{a}{4} = 1$

$\therefore$ $\frac{5}{2}a = 1$ or $a = \frac{2}{5}$

$$p(x > 4) = p(x = 5) = \frac{9}{4} = \frac{1}{10}$$

$$p(x \leq 4) = 0 + a + \frac{a}{2} + \frac{a}{2} + \frac{a}{4} + \frac{9a}{4} = \frac{9}{10}$$

Example 6.9: A fair coin is tossed 400 times. Find the mean number of heads and the corresponding standard deviation.

Solution:

This is a case of binomial distribution with $p = q = \frac{1}{2}$, $n = 400$.

The mean number of heads is given by $\mu = np = 400 \times \frac{1}{2} = 200$.

and S. D. $\sigma = \sqrt{npq} = \sqrt{400 \times \frac{1}{2} \times \frac{1}{2}} = 10$

Example 6.10: A manager has thought of 4 planning strategies each of which has an equal chance of being successful. What is the probability that at least one of his strategies will work if he tries them in 4 situations? Here $p = \frac{1}{4}, q = \frac{3}{4}$.

NOTES

Solution:

The probability that none of the strategies will work is given by,

$$p(0) = {}^4C_0 \left(\frac{1}{4}\right)^0 \left(\frac{3}{4}\right)^4 = \left(\frac{3}{4}\right)^4$$

The probability that at least one will work is given by $1 - \left(\frac{3}{4}\right)^4 = \frac{175}{256}$.

Example 6.11: For the Poisson distribution, write the probabilities of 0, 1, 2, ... successes.

Solution:

x	$p(x) = e^{-m} \frac{m^x}{x!}$
0	$p(0) = e^{-m} m^0 / 0!$
1	$p(1) = e^{-m} \frac{m}{1!} = p(0).m$
2	$e^{-m} \frac{m^2}{2!} = p(2) = p(1).\frac{m}{2}$
3	$e^{-m} \frac{m^3}{3!} = p(3) = p(2).\frac{m}{3}$
⋮	

and so on.

Total of all probabilities $\Sigma p(x) = 1$.

Example 6.12: What are the raw moments of Poisson distribution?

Solution:

First raw moment $\mu'_1 = m$

Second raw moment $\mu'_2 = m^2 + m$

Third raw moment $\mu'_3 = m^3 + 3m^2 + m$

(v) Continuous Probability Distributions

When a random variate can take any value in the given interval $a \le x \le b$, it is a continuous variate and its distribution is a continuous probability distribution.

NOTES

Theoretical distributions are often continuous. They are useful in practice because they are convenient to handle mathematically. They can serve as good approximations to discrete distributions.

The range of the variate may be finite or infinite.

Acontinuous random variable can take all values in a given interval. Acontinuous probability distribution is represented by a smooth curve.

The total area under the curve for a probability distribution is necessarily unity. The curve is always above the *x* axis because the area under the curve for any interval represents probability and probabilities cannot be negative.

If *X* is a continous variable, the probability of *X* falling in an interval with end points z_1, z_2 may be written $p(z_1 \leq X \leq z_2)$.

This probability corresponds to the shaded area under the curve in Figure 6.5.

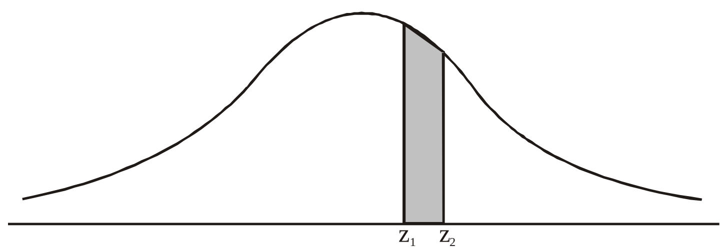


Fig. 6.5 *Continuous Probability Distribution*

A function is a probability density function if,

$\int_{-\infty}^{\infty} p(x)dx = 1, p(x) \geq 0, -\infty < x < \infty,$ i.e., the area under the curve $p(x)$ is 1 and the probability of *x* lying between two values *a*, *b*, i.e., $p(a < x < b)$ is positive. The most prominent example of a continuous probability function is the normal distribution.

Cumulative Probability Function (CPF)

The Cumulative Probability Function (CPF) shows the probability that *x* takes a value less than or equal to, say, *z* and corresponds to the area under the curve up to *z:*

$$p(x \leq z) = \int_{-\infty}^{z} p(x)dx$$

This is denoted by $F(x)$.

1.4.3 Extension to Bivariate Case: Elementary Concepts

If in a bivariate distribution the data is quite large, then they may be summed up in the form of a two-way table. In this for each variable, the values are grouped into different classes (not necessary same for both the variables), keeping in view the same considerations as in the case of univariate distribution. In other words, a bivariate frequency distribution presents in a table pairs of values of two variables and their frequencies.

For example, if there is m classes for the *X* – variable series and n classes for the *Y* – variable series then there will be $m \times n$ cells in the two-way table. By going through the different pairs of the values (*x*, *y*) and using tally marks, we can find the frequency for each cell and thus get the so called bivariate frequency table.

Check Your Progress

4. What is Bayes' theorem?
5. What is joint probability?
6. When a distribution is said to be symmetrical?
7. What is continuous probability distribution?

NOTES

Table 6.3 *Bivariate Frequency Table*

x Series → / *Y* Series ↓	Classes Mid Points x_1 x_1 ....x .... x_m	Total of Frequencies of *y*
Classes / Mid Points y_1 y_2 y . . . . y_n	$f(x, y)$	f_y
Total of Frequencies of *x*	f_x	Total $f_x = f_y = \text{N}$

Here, $f(x,y)$ is the frequency of the pair (x, y). The formula for computing the correlation coefficient between x and y for the bivariate frequency table is,

$$r = \frac{N\Sigma xyf\ (x, y) - (\Sigma xfx)(\Sigma yfy)}{\sqrt{\left[N\Sigma x^2 f_x - (\Sigma xfx)^2\right] \times \left[N\Sigma y^2 f_y - (\Sigma yfy)^2\right]}}$$

where, N is the total frequency.

1.5 SUMMARY

- The probability theory helps a decision-maker to analyse a situation and decide accordingly.
- Bayes' theorem makes use of conditional probability formula where the condition can be described in terms of the additional information which would result in the revised probability of the outcome of an event.
- A random variable takes on different values as a result of the outcomes of a random experiment.
- When a random variate can take any value in the given interval $a \le x \le b$, it is a continuous variate and its distribution is a continuous probability distribution.

1.6 KEY TERMS

- **Axiomatic probability theory:** The most general approach to probability, and is used for more difficult problems in probability
- **Event:** An outcome or a set of outcomes of an activity or a result of a trial
- **Random variable:** A variable that takes on different values as a result of the outcomes of a random experiment

1.7 ANSWERS TO 'CHECK YOUR PROGRESS'

NOTES

1. Different types of probability theories are:
   (i) Axiomatic Probability Theory
   (ii) Classical Theory of Probability
   (iii) Empirical Probability Theory
2. The classical theory of probability is based on the number of favourable outcomes and the number of total outcomes.
3. The Law of Large Numbers (LLN) states that as the number of trials of an experiment increases, the empirical probability approaches the theoretical probability. Hence, if we roll a die a number of times, each number would come up approximately 1/6 of the time.
4. Bayes' theorem makes use of conditional probability formula where the condition can be described in terms of the additional information which would result in the revised probability of the outcome of an event.
5. The product of prior probability and conditional probability for each state of nature is called joint probability.
6. If $p = 0.5$, the distribution is symmetrical.
7. When a random variate can take any value in the given interval $a \le x \le b$, it is a continuous variate and its distribution is a continuous probability distribution.

1.8 QUESTIONS AND EXERCISES

Short-Answer Questions

1. List the various types of events.
2. What are independent events?
3. What are the rules of assigning probability?
4. What are the three techniques of assigning probability?
5. What is Bernoulli trials?
6. What are the significant characteristics of binomial distribution?
7. What is CPF?

Long-Answer Questions

1. Explain the various theories of probability with the help of example.
2. Discuss the law of addition and law of multiplication with the help of example.
3. Describe Bayes' theorem with the help of an example.
4. Analyse the types of discrete distributions.

1.9 FURTHER READING

NOTES

Allen, R.G.D. 2008. *Mathematical Analysis For Economists*. London: Macmillan and Co., Limited.

Chiang, Alpha C. and Kevin Wainwright. 2005. *Fundamental Methods of Mathematical Economics*, 4 edition. New York: McGraw-Hill Higher Education.

Yamane, Taro. 2012. *Mathematics For Economists: An Elementary Survey*. USA: Literary Licensing.

Baumol, William J. 1977. *Economic Theory and Operations Analysis*, 4th revised edition. New Jersey: Prentice Hall.

Hadley, G. 1961. *Linear Algebra*, 1st edition. Boston: Addison Wesley.

Vatsa , B.S. and Suchi Vatsa. 2012. *Theory of Matrices*, 3rd edition. London: New Academic Science Ltd.

Madnani, B C and G M Mehta. 2007. *Mathematics for Economists*. New Delhi: Sultan Chand & Sons.

Henderson, R E and J M Quandt. 1958. *Microeconomic Theory: A Mathematical Approach*. New York: McGraw-Hill.

Nagar, A.L. and R.K.Das. 1997. *Basic Statistics*, 2nd edition. United Kingdom: Oxford University Press.

Gupta, S.C. 2014. *Fundamentals of Mathematical Statistics*. New Delhi: Sultan Chand & Sons.

M. K. Gupta, A. M. Gun and B. Dasgupta. 2008. *Fundamentals of Statistics*. West Bengal: World Press Pvt. Ltd.

Saxena, H.C. and J. N. Kapur. 1960. *Mathematical Statistics*, 1st edition. New Delhi: S. Chand Publishing.

Hogg, Robert V., Joeseph McKean and Allen T Craig. *Introduction to Mathematical Statistics*, 7th edition. New Jersey: Pearson.

UNIT II PROBABILITY DISTRIBUTION

NOTES

Structure

2.0 INTRODUCTION

In this unit, you will learn about expectation and its properties. Expected value of X is the weighted average of the possible values that X can take. You will be familiarized with mean, variance and moments in terms of expectation. Also, the unit will explain the process of summing two random variables, you will learn about variance and standard deviation of random variable and finally about moments generating functions. This unit will also discuss standard distribution and statistical inference in detail.

You will learn that a random variable is a function that associates a unique numerical value with every outcome of an experiment. The value of the random variable will vary from trial to trial as the experiment is repeated. There are two types of random variables—discrete and continuous. The probability distribution of a discrete random variable is a list of probabilities associated with each of its possible values. It is also sometimes called the probability function or probability mass function. The probability density function of a continuous random variable is a function which can be integrated to obtain the probability that the random variable takes a value in a given interval. Binomial distribution is used in finite sampling problems where each observation has one of two possible outcomes

NOTES

(‘success’ or ‘failure’). Poisson distribution is used for modelling rates of occurrence. Exponential distribution is used to describe units that have a constant failure rate. The term ‘normal distribution’ refers to a particular way in which observations will tend to pile up around a particular value rather than be spread evenly across a range of values, i.e., the ‘Central Limit Theorem’. It is generally most applicable to continuous data and is intrinsically associated with parametric statistics (for example, ANOVA, *t*-test, regression analysis). Graphically, the normal distribution is best described by a bell-shaped curve. This curve is described in terms of the point at which its height is maximum, i.e., its mean and its width or standard deviation.

2.1 UNIT OBJECTIVES

After going through this unit, you will be able to:

- Understand about expectation and its properties
- Discuss about mean, variance and moments in terms of expectation
- Understand moments generating functions
- Discuss about standard distribution and statistical inference
- Understand the basic concept of probability distribution
- Explain the types of probability distribution
- Describe the binomial distribution based on Bernoulli process
- Describe the significance of Poisson distribution
- Explain the basic theory, characteristics and family of normal distributions
- Measure the area under the normal curve
- Analyse Poisson distribution as an approximation of binomial distribution
- Explain the beta and gamma distribution

2.2 EXPECTATION AND ITS PROPERTIES

The expected value (or mean) of X is the weighted average of the possible values that X can take. Here, X is a discrete random variable and each value is being weighted according to the probability of the possibility of occurrence of the event. The expected value of X is usually written as $E(X)$ or μ.

$$E(X) = \Sigma \times P(X = x)$$

Hence, the expected value is the sum of each of the possible outcomes and the probability of the outcome occurring.

Therefore, the expectation is the outcome you expect of an experiment.

Let us consider the Example 7.1.

Example 7.1: What is the expected value when we roll a fair die?

Solution: There are six possible outcomes 1, 2, 3, 4, 5, 6. Each one of these has a probability of 1/6 of occurring. Let X be the outcome of the experiment.

Then,

$P(X = 1) = 1/6$ (this shows that the probability that the outcome of the experiment is 1 is 1/6)

$P(X = 2) = 1/6$ (the probability that you throw a 2 is 1/6)

$P(X = 3) = 1/6$ (the probability that you throw a 3 is 1/6)

$P(X = 4) = 1/6$ (the probability that you throw a 4 is 1/6)

$P(X = 5) = 1/6$ (the probability that you throw a 5 is 1/6)

$P(X = 6) = 1/6$ (the probability that you throw a 6 is 1/6)

$E(X) = 1 \times P(X = 1) + 2 \times P(X = 2) + 3 \times P(X = 3) + 4 \times P(X=4) + 5 \times P(X=5) + 6 \times P(X = 6)$

Therefore,

$$E(X) = 1/6 + 2/6 + 3/6 + 4/6 + 5/6 + 6/6 = 7/2 \text{ or } 3.5$$

Hence, the expectation is 3.5, which is also the halfway between the possible values the die can take, and so this is what you should have expected.

Expected Value of a Function of *X*

To find $E[f(X)]$, where $f(X)$ is a function of X, we use the formula,

$$E[f(X)] = \Sigma f(x)P(X = x)$$

Let us consider Example 7.1 of die, and calculate $E(X^2)$

Using the notation above, $f(x) = x^2$

$f(1) = 1, f(2) = 4, f(3) = 9, f(4) = 16, f(5) = 25, f(6) = 36$

$P(X = 1) = 1/6, P(X = 2) = 1/6$, etc.

Hence, $E(X^2) = 1/6 + 4/6 + 9/6 + 16/6 + 25/6 + 36/6 = 91/6 = 15.167$

The expected value of a constant is just the constant, as for example $E(1) = 1$. Multiplying a random variable by a constant multiplies the expected value by that constant.

Therefore, $E[2X] = 2E[X]$

An important formula, where a and b are constants is,

$$E[aX + b] = aE[X] + b$$

Hence, we can say that the expectation is a linear operator.

Variance

The variance of a random variable tells us something about the spread of the possible values of the variable. For a discrete random variable X, the variance of X is written as Var(X).

$$\text{Var}(X) = E[(X - \mu)^2]$$

Where, μ is the expected value $E(X)$

This can also be written as,

$$\text{Var}(X) = E(X^2) - \mu^2$$

The standard deviation of X is the square root of Var(X).

***Note*:** The variance does not behave in the same way as expectation, when we multiply and add constants to random variables.

$$\text{Var}[aX + b] = a^2\text{Var}(X)$$

Because, $\text{Var}[aX + b] = E[\ (aX + b)^2\ ] - (E\ [aX + b])^2$

$$= E[\ a^2X^2 + 2abX + b^2] - (aE(X) + b)^2$$
$$= a^2E(X^2) + 2abE(X) + b^2 - a^2E^2(X) - 2abE(X) - b^2$$
$$= a^2E(X^2) - a^2E^2(X) = a^2\text{Var}(X)$$

Expectation (Conditional)

The expectation of a random variable X with Probability Density Function (PDF) $p(x)$ is theoretically defined as:

$$E[X] = \int xp(x)dx$$

If we consider two random variables X and Y (not necessarily independent), then their combined behaviour is described by their joint probability density function $p(x, y)$ and is defined as,

The marginal probability density of X is defined as,

$$p_X(x) = \int p(x, y)dy$$

For any fixed value y of Y, the distribution of X is the conditional distribution of X, where $Y = y$, and it is denoted by $p(x , y)$.

Expectation (Iterated)

The expectation of the random variable is expressed as,

$$E[X] = E[E[X \mid Y]]$$

This expression is known as the 'Theorem of Iterated Expectation' or 'Theorem of Double Expectation'. Symbolically, it can be expressed as,

(i) For the discrete case,

$$E[X] = \Sigma_y\ E[X \mid Y = y].P\{Y = y\}$$

(ii) For the continuous case,

$$E[X] = \int_{-\infty}^{+\infty} E[X|Y = y].\ f(y).dy$$

Expectation: Continuous Variables

If x is a continuous random variable we define that,

$$E(x) =$$

The expectation of a function $h(x)$ is,

$$Eh(x) = \int_{-\infty}^{\infty} h(x)\ P(x)dx$$

NOTES

The rth moment about the mean is,

$$E(x-\mu)^r = \int_{-\infty}^{\infty} (x-\mu)^r \ P(x)dx$$

Consider the following examples.

Example 7.2: A newspaper seller earns ` 100 a day if there is suspense in the news. He loses ` 10 a day if it is an eventless newspaper. What is the expectation of his earnings if the probability of suspense news is 0.4?

Solution:

$$E(x) = p_1x_1 + p_2x_2$$
$$= 0.4 \times 100 - 0.6 \times 10$$
$$= 40 - 6 = 34$$

Example 7.3: A player tossing three coins, earns ` 10 for 3 heads, ` 6 for 2 heads and ` 1 for 1 head. He loses ` 25, if 3 tails appear. Find his expectation.

Solution:

$$p(HHH) = \frac{1}{2}\cdot\frac{1}{2}\cdot\frac{1}{2} = \frac{1}{8} = p_1 \text{ say} \qquad \text{3 heads}$$

$$p(HHT) = {}^3C_2 \cdot \frac{1}{2}\cdot\frac{1}{2}\cdot\frac{1}{2} = \frac{3}{8} = p_2 \text{ say} \qquad \text{2 heads, 1 tail}$$

$$p(HTT) = {}^3C_1 \cdot \frac{1}{2}\cdot\frac{1}{2}\cdot\frac{1}{2} = \frac{3}{8} = p_3 \text{ say} \qquad \text{1 head, 2 tails}$$

$$p(TTT) = \frac{1}{2}\cdot\frac{1}{2}\cdot\frac{1}{2} = \frac{3}{8} = p_4 \text{ say} \qquad \text{3 tails}$$

$\frac{9}{8}$ = ` 1.125

$$E(x) = p_1x_1 + p_2x_2 + p_3x_3 - p_4x_4$$

$$= \frac{1}{8}\times 10 + \frac{3}{8}\times 6 + \frac{3}{8}\times 2 - \frac{1}{8}\times 25$$

=

Example 7.4: Calculate the Standard Deviation (S.D.) when x takes the values 0, 1, 2 and 9 with probability, 0.4, 0 – 0.1.

Solution:

x takes the values 0, 1, 2, 9 with probability 0.4, 0.2, 0.3, 0.1.

$$\mu = E(x) = \Sigma(x_i p_i) = 0 \times 0.4 + 1^2 \times 0.2 + 3 \times 0.3 + 9 \times 0.1 = 2.0$$

$$E(x^2) = \Sigma x_i^2 p_i = 0^2 \times 0.4 + 1^2 \times 0.2 + 3 \times 0.3 + 9^2 \times 0.1 = 11.0$$

$$V(x) = E(x^2) - \mu^2 = 11 - 2 = 9$$

$$\text{S.D.}(x) = \sqrt{9} = 3$$

Example 7.5: The purchase of some shares can give a profit of ` 400 with probability 1/100 and ` 300 with probability 1/20. Comment on a fair price of the share.

Solution:

Expected value $E(x) = \Sigma x_i p_i =$

NOTES

2.2.1 Mean, Variance and Moments in Terms of Expectation

Mean of random variable is the sum of the values of the random variable weighted by the probability that the random variable will take on the value. In other words, it is the sum of the product of the different values of the random variable and their respective probabilities. Symbolically, we write the mean of a random variable, say *X*, as . The expected value of the random variable is the average value that would occur if we have to average an infinite number of outcomes of the random variable. In other words, it is the average value of the random variable in the long run. The expected value of a random variable is calculated by weighting each value of a random variable by its probability and summing over all values. The symbol for the expected value of a random variable is *E* (*X*). Mathematically, we can write the mean and the expected value of a random variable, *X*,

and,

$$E \quad X \quad \overset{n}{\underset{i\ 1}{}} \ X_i \ .pr. \ X_i$$

Thus, the mean and expected value of a random variable are conceptually and numerically the same, but usually denoted by different symbols and as such the two symbols, viz., $\overline{X}$ and *E* (*X*) are completely interchangeable. We can, therefore, express the two as,

$$E \quad X \quad \overset{n}{\underset{i\ 1}{}} \ X_i \ .pr. \ X_i \quad \overline{X}$$

Where, X_i is the *i*th value *X* can take.

Sum of Random Variables

If we are given the means or the expected values of different random variables, say *X*, *Y*, and *Z* to obtain the mean of the random variable (*X* + *Y* + *Z*), then it can be obtained as,

$$E \quad X \quad Y \quad Z \quad E \quad X \quad E \quad Y \quad E \quad Z \quad \overline{X} \quad \overline{Y} \quad \overline{Z}$$

Similarly, the expectation of a constant time random variable is the expected value of the random variable. Symbolically, we can write this as,

Where *cX* is the constant time random variable.

Variance and Standard Deviation of Random Variable

The mean or the expected value of random variable may not be adequate enough at times to study the problem as to how random variable actually behaves and we may as well be interested in knowing something about how the values of random variable are dispersed about the mean. In other words, we want to measure the dispersion of

NOTES

random variable (X) about its expected value, i.e., E(X). The variance and the standard deviation provide measures of this dispersion.

The variance of random variable is defined as the sum of the squared deviations of the values of random variable from the expected value weighted by their probability. Mathematically, we can write it as,

$$\text{Var}\, X = \sigma_X^2 = \sum_{i=1}^{n} \left[X_i - E(X)\right]^2 . pr.(X_i)$$

Alternatively, it can also be written as,

$$\text{Var}\, X = \sigma_X^2 = \sum X_i^2\, pr.(X_i) - \left[E(X)\right]^2$$

Where, $E(X)$ is the expected value of random variable.

X_i is the ith value of random variable.

$pr.(X_i)$ is the probability of the ith value.

The standard deviation of random variable is the square root of the variance of random variable and is denoted as,

$$\sqrt{\sigma_X^2} = \sigma_X$$

The variance of a constant time random variable is the constant squared times the variance of random variable. This can be symbolically written as,

$$\text{Var}\,(cX) = c^2\, \text{Var}\,(X)$$

The variance of a sum of independent random variables equals the sum of the variances. Thus,

$$\text{Var}\,(X + Y + Z) = \text{Var}(X) + \text{Var}(Y) + \text{Var}(Z)$$

If X, Y and Z are independent of each other.

The following examples will illustrate the method of calculation of these measures of a random variable.

Example 7.6: Calculate the mean, variance and standard deviation of random variable sales from the following information provided by a sales manager of a certain business unit for a new product:

Monthly Sales (in units)	*Probability*
50	0.10
100	0.30
150	0.30
200	0.15
250	0.10
300	0.05

Solution: The given information may be developed as shown in the following table for calculating mean, variance and the standard deviation for random variable sales:

NOTES

Monthly Sales (in units)[1] X_i		Probability $pr\ (X_i)$	$(X_i)\ pr\ (X_i)$	$(X_i - E\ (X))^2$	$X_i - E\ (X)^2\ pr\ (X_i)$
X_1	50	0.10	5.00	$(50 - 150)^2 = 10000$	1000.00
X_2	100	0.30	30.00	$(100 - 150)^2 = 2500$	750.00
X_3	150	0.30	45.00	$(150 - 150)^2 = 0$	0.00
X_4	200	0.15	30.00	$(200 - 150)^2 = 2500$	375.00
X_5	250	0.10	25.00	$(250 - 150)^2 = 10000$	1000.00
X_6	300	0.5	15.00	$(300 - 150)^2 = 22500$	1125.00
			$\sum(X_i)\ pr\ (X_i) = 150.00$		$\sum[X_i - E\ (X)^2]\ pr\ (X_i) = 4250.00$

Mean of random variable sales = $\bar{X}$

or, $E\ (X) = \sum(X_i).pr(X_i) = 150$

Variance of random variable sales,

or,

Standard deviation of random variable sales,

or, $_X \quad \sqrt{^2_X} \quad \sqrt{4250} \quad 65.2$ approx.

The mean value calculated above indicates that in the long run the average sales will be 150 units per month. The variance and the standard deviations measure the variation or dispersion of random variable values about the mean or the expected value of random variable.

Example 7.7: Given are the mean values of four different random variables viz., *A*, *B*, *C* and *D*.

$$\bar{A} \quad 20, \quad \bar{B} \quad 40, \quad \bar{C} \quad 10, \quad \bar{D} \quad 5$$

Find the mean value of the random variable $(A + B + C + D)$.

Solution:

$E(A + B + C + D) = E(A) + E(B) + E(C) + E(D)$

$=$

$= 20 + 40 + 10 + 5$

$= 75$

Hence, the mean value of random variable $(A + B + C + D)$ is 75.

1. Computations can be made easier if we take 50 units of sales as one unit in the given example, such as 100 as 2, 200 as 4 and so on.

2.2.2 Moment Generating Functions

According to probability theory, moment generating function generates the moments for the probability distribution of a random variable *X*, and can be defined as,

$$M_X(t) = E(e^{tX}),$$

When the moment generating function exists with an interval $t = 0$, the nth moment becomes,

$$E(X^n) = M_X^{(n)}(0) = [d^n M_X(t) / dt^n]_{t=0}$$

The moment generating function, for probability distribution condition being continuous or not, can also be given by Riemann-Stieltjes integral,

Where *F* is the cumulative distribution function.

The probability density function $f(x)$, for *X* having continuous moment generating function becomes,

$$M_X(t) = \int_{-\infty}^{\infty} e^{tX} f(x)\, dx$$

$$= \int_{-\infty}^{\infty} (1 + tx + t^2x^2/2! + \ldots) f(x)\, dx$$

***Note*:** The moment generating function of *X* always exists, when the exponential function is positive and is either a real number or a positive infinity.

$$M_X(t) = \int_{-\infty}^{\infty} e^{tx}\, dF(x)$$

$$M_X(t) = \int_S e^{tx} f(x)\, dx$$

1. Prove that when *X* shows a discrete distribution having density function *f*, then,

$$M_X(t) = \sum_{x \in S} e^{tx} f(x)$$

2. When *X* is continuous with density function *f*, then

3. Consider that *X* and *Y* are independent. Show that,

$$M_{X+Y}(t) = M_X(t)\, M_Y(t)$$

2.3 STANDARD DISTRIBUTION

The standard distribution is a special case of the normal distribution. It is the distribution that occurs when a normal random variable has a mean of zero and a standard deviation of one.

The normal random variable of a standard normal distribution is called a standard score or a *z*-score. Every normal random variable *X* can be transformed into a z-score through the equation,

$$z = (X - \mu) / \sigma$$

NOTES

Where X is a normal random variable, μ is the mean of X, and σ is the standard deviation of X.

For example, if a person scored a 70 on a test with a mean of 50 and a standard deviation of 10, then they scored 2 standard deviations above the mean. Converting the test scores to z scores, an X of 70 would be,

$$z = \frac{70 - 50}{10} = 2$$

So, a z-score of 2 means the original score was 2 standard deviations above the mean.

Standard Normal Distribution Table

A standard normal distribution table shows a cumulative probability associated with a particular z-score. Table rows show the whole number and tenths place of the z-score. Table columns show the hundredths place. The cumulative probability (often from minus infinity to the z-score) appears in the cell of the table. This is further discussed in Unit 7.

2.4 STATISTICAL INFERENCE

Statistical inference refers to drawing conclusions based on data that is subjected to random variation; for example, sampling variation or observational errors. The terms statistical inference, statistical induction and inferential statistics are used to describe systems of procedures that can be used to draw conclusions from datasets arising from systems affected by random variation.

There are many contexts in which inference is desirable, and there are also various approaches of performing inferences. One of the most important contexts is parametric models. For example, if you have noisy (x, y) data that you think follow the pattern $y = \beta_0 + \beta_1 x$ + error, then you can estimate β_0, β_1, and the magnitude of the error.

The fundamental requirements of such set of procedures for inference are that they must be common so that it can be applied on a range of conditions and produce a logical and reasonable conclusion whenever applied to well-defined and simple situations. The result of this procedure, when used in analysis of statistical data is generally an estimate or a set of estimates of one or more parameters that describe the problem along with some indication of uncertainty with which the values are estimated. This technique is different from descriptive statistics in respect that descriptive statistics is just a straightforward presentation of facts, in which decisions are made by influence of data analyst; whereas there is no influence of analyst in statistical inference.

The method of statistical inference is generally used for point estimation, interval estimation, hypothesis testing or statistical significance testing and prediction of a random process.

Statistical inference is generally distinguished from descriptive statistics. In simple terms, descriptive statistics can be thought of as being just a straightforward presentation of facts, in which modeling decisions made by a data analyst have had minimal influence. Any statistical inference requires some assumptions. A statistical model is a set of assumptions concerning the generation of observed data and similar data. A complete statistical analysis will nearly always include both descriptive statistics and statistical inference, and will often progress in a series of steps where the emphasis moves gradually from description to inference.

NOTES

2.5 BINOMIAL DISTRIBUTION

Binomial distribution (or the Binomial probability distribution) is a widely used probability distribution concerned with a discrete random variable and as such is an example of a discrete probability distribution. The binomial distribution describes discrete data resulting from what is often called the Bernoulli process. The tossing of a fair coin a fixed number of times is a Bernoulli process and the outcome of such tosses can be represented by the binomial distribution. The name of Swiss mathematician Jacob Bernoulli is associated with this distribution. This distribution applies in situations where there are repeated trials of any experiment for which only one of two mutually exclusive outcomes (often denoted as 'success' and 'failure') can result on each trial.

2.5.1 Bernoulli Process

Binomial distribution is considered appropriate in a Bernoulli process which has the following characteristics:

2.5.1.1 **Dichotomy:** This means that each trial has only two mutually exclusive possible outcomes, e.g., 'Success' or 'failure', 'Yes' or 'No', 'Heads' or 'Tails' and the like.

2.5.1.2 **Stability:** This means that the probability of the outcome of any trial is known (or given) and remains fixed over time, i.e., remains the same for all the trials.

2.5.1.3 **Independence:** This means that the trials are statistically independent, i.e., to say the happening of an outcome or event in any particular trial is independent of its happening in any other trial or trials.

2.5.2 Probability Function of Binomial Distribution

The random variable, say *X*, in the binomial distribution is the number of 'successes' in *n* trials. The probability function of the binomial distribution is written as,

$$f(X = r) = {}^nC_r p^r q^{n-r}$$

$$r = 0, 1, 2 \ldots n$$

Where, n = Numbers of trials.

p = Probability of success in a single trial.

$q = (1 - p)$ = Probability of 'failure' in a single trial.

r = Number of successes in 'n' trials.

2.5.3 Parameters of Binomial Distribution

Binomial distribution depends upon the values of *p* and *n* which in fact are its parameters. Knowledge of *p* truly defines the probability of *X* since *n* is known by definition of the problem. The probability of the happening of exactly *r* events in *n* trials can be found out using the previously stated binomial function.

The value of *p* also determines the general appearance of the binomial distribution, if shown graphically. In this context the usual generalizations are as follows:

2.5.3.1 When *p* is small (say 0.1), the binomial distribution is skewed to the right, i.e., the graph takes the form shown in Figure 7.1.

Check Your Progress

1. What is expected value?
2. What is the mean of random variable?
3. How is expected value of a random variable calculated?
4. What is *z*-score?
5. Define the term statistical inference.

NOTES

Fig. 7.1 Graph When p<0.1

2.5.3.2 When p is equal to 0.5, the binomial distribution is symmetrical and the graphtakes the form as shown in Figure 7.2.

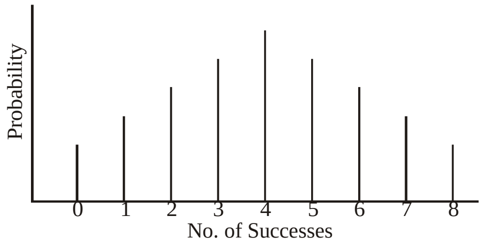


Fig. 7.2 *Graph When p = 0.5*

2.5.3.3 When p is larger than 0.5, the binomial distribution is skewed to the left and thegraph takes the form as shown in Figure 7.3.

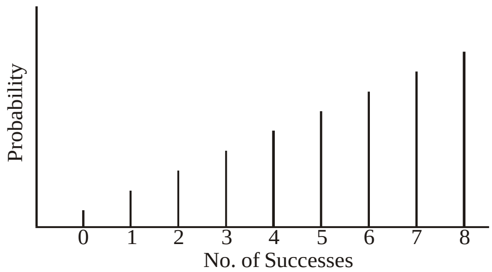


Fig. 7.3 *Graph When p>0.5*

If, however, 'p' stays constant and 'n' increases, then as 'n' increases the vertical lines become not only numerous, but also tend to bunch up together to form a bell shape, i.e., the binomial distribution tends to become symmetrical and the graph takes the shape as shown in Figure 7.4.

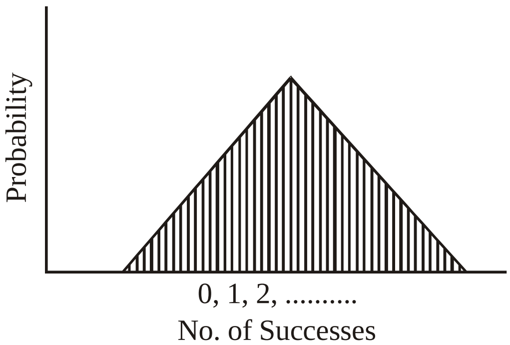


Fig. 7.4 *Graph When p is Constant*

2.5.4 Important Measures of Binomial Distribution

The ex_pected value of random variable [i.e., $E(X)$] or mean of random variable (i.e., $\bar{X}$) of the binomial distribution is equal to $n.p$ and the variance of random variable is equal to $n.\ p.\ q$ or $n.\ p.\ (1 - p)$. Accordingly, the standard deviation of binomial distribution is equal to   The other important measures relating to binomial distribution are as,

Skewness =

Kurtosis =

2.5.5 When to Use Binomial Distribution

The use of binomial distribution is most appropriate in situations fulfilling the conditions outlined in Section 7.2.4. Two such situations, for example, can be described as follows.

- When we have to find the probability of 6 heads in 10 throws of a fair coin.
- When we have to find the probability that 3 out of 10 items produced by a machine, which produces 8 per cent defective items on an average, will be defective.

Example 7.1: A fair coin is thrown 10 times. The random variable X is the number of head(s) coming upwards. Using the binomial probability function, find the probabilities of all possible values which X can take and then verify that binomial distribution has a mean: $\bar{X} = n.p.$ and variance: $= n.p.q.$

Solution: Since the coin is fair and so, when thrown, can come either with head upward or tail upward. Hence, p (head)   and q (no head)   The required probability function is,

$1 - 2p$ $6n.p.q\ 6p$ $6q^2$ $\sqrt{\ }$ 3 $\sqrt{2}$ $n.p.q$ $n.p.q$

$$f(X = r) = {}^nC_r\ p^r q^{n-r}$$

$$r = 0, 1, 2 \ldots 10$$

The following table of binomial probability distribution is constructed using this function:

X_i (*Number of Heads*)	*Probability* pr_i	$X_i\ pr_i$	$(X_i - \bar{X})$	$(X_i - \bar{X})^2$	$(X_i - \bar{X})^2.p_i$
0	$^{10}C_0\ p^0\ q^{10}$ = 1/1024	0/1024	–5	25	25/1024
1	$^{10}C_1\ p^1\ q^9$ = 10/1024	10/1024	–4	16	160/1024
2	$^{10}C_2\ p^2\ q^8$ = 45/1024	90/1024	–3	9	405/1024
3	$^{10}C_3\ p^3\ q^7$ = 120/1024	360/1024	–2	4	480/1024
4	$^{10}C_4\ p^4\ q^6$ = 210/1024	840/1024	–1	1	210/1024
5	$^{10}C_5\ p^5\ q^5$ = 252/1024	1260/1024	0	0	0/1024
6	$^{10}C_6\ p^6\ q^4$ = 210/1024	1260/1024	1	1	210/1024
7	$^{10}C_7\ p^7\ q^3$ = 120/1024	840/1024	2	4	480/1024
8	$^{10}C_8\ p^8\ q^2$ = 45/1024	360/1024	3	9	405/1024
9	$^{10}C_9\ p^9\ q^1$ = 10/1024	90/1024	4	16	160/1024
10	$^{10}C_{10}\ p^{10}\ q^0$ = 1/1024	10/1024	5	25	25/1024
		ΣX = 5120/1024 $\bar{X}$ = 5			Variance = σ^2 = $\Sigma (X_i - \bar{X})^2.pr_i$ = 2560/1024 = 2.5

NOTES

The mean of the binomial distribution[2] is given by *n. p.* = 10 × = 5 and the

variance of this distribution is equal to *n. p. q.* = 10 × × = 2.5.

These values are exactly the same as we have found them in the table. Hence, these values stand verified with the calculated values of the two measures as shown in the table.

2.6 POISSON DISTRIBUTION

Poisson distribution is also a discrete probability distribution which is associated with the name of a Frenchman, Simeon Denis Poisson who developed this distribution. Poisson distribution is frequently used in context of Operations Research and for this reason has a great significance for management people. This distribution plays an important role in queuing theory, inventory control problems and also in risk models.

Unlike binomial distribution, Poisson distribution cannot be deducted on purely theoretical grounds based on the conditions of the experiment. In fact, it must be based on experience, i.e., on the empirical results of past experiments relating to the problem under study. Poisson distribution is appropriate especially when the probability of happening of an event is very small (so that q or $(1-p)$ is almost equal to unity) and n is very large, such that the average of series (viz., *n. p.*) is a finite number. Experience has shown that this distribution is good for calculating the probabilities associated with X occurrences in a given time period or specified area.

The random variable of interest in Poisson distribution is the number of occurrences of a given event during a given interval (interval may be time, distance, area, etc.). We use capital X to represent the discrete random variable and lower case x to represent a specific value that capital X can take. The probability function of this distribution is generally written as,

$$x = 0, 1, 2\ldots$$

Where, λ = Average number of occurrences per specified interval[3]. In other words, it is the mean of the distribution.

e = 2.7183 being the basis of natural logarithms.

x = Number of occurrences of a given event.

2. The value of the binomial probability function for various values of n and p are also available in tables (known as binomial tables) which can be used for the purpose to ease calculation work. The tables are of considerable help particularly when n is large.

3. For Binomial distribution we had stated that mean = *n. p.*

$\therefore$ Mean for Poisson distribution (or λ) = *n. p.*

$\therefore$ $p = \lambda/n$

Hence, mean =

NOTES

Poisson Process

The characteristics of Poisson process are as follows:

(i) Concerning a given random variable, the mean relating to a given interval can be estimated on the basis of past data concerning the variable under study.

(ii) If we divide the given interval into very very small intervals we will find the following:

(a) The probability that exactly one event will happen during the very very small interval is a very small number and is constant for every other very small interval.

(b) The probability that two or more events will happen within a very small interval is so small that we can assign it a zero value.

(c) The event that happens in a given very small interval is independent, when the very small interval falls during a given interval.

(d) The number of events in any small interval is not dependent on the number of events in any other small interval.

Parameter and Important Measures of Poisson Distribution

Poisson distribution depends upon the value of λ, the average number of occurrences per specified interval which is its only parameter. The probability of exactly x occurrences can be found out using Poisson probability function[4]. The expected value or the mean of Poisson random variable is λ and its variance is also λ[5]. The standard deviation of Poisson distribution is, .

Underlying the Poisson model is the assumption that if there are on the average λ occurrences per interval t, then there are on the average $k\,\lambda$ occurrences per interval kt. For example, if the number of arrivals at a service counted in a given hour, has a Poisson distribution with $\lambda = 4$, then y, the number of arrivals at a service counter in a given 6 hour day, has the Poisson distribution $\lambda = 24$, i.e., 6×4.

$\sqrt{\lambda}$

x

When to Use Poisson Distribution

The use of Poisson distribution is resorted to in cases when we do not know the value of '*n*' or when '*n*' cannot be estimated with any degree of accuracy. In fact, in certain cases it does not make any sense in asking the value of '*n*'. For example, the goals scored by one team in a football match are given, it cannot be stated how many goals could not be scored. Similarly, if one watches carefully one may find out how many times the lightning flashed, but it is not possible to state how many times it did not flash. It is in such cases we use Poisson distribution. The number of deaths per day in a district in one year due to a disease, the number of scooters passing through a road per minute during a certain part of the day for a few months, the number of printing mistakes per page in a book containing many pages, are a few other examples where Poisson probability distribution is generally used.

---

4. There are tables which give the $e^{-\lambda}$ values. These tables also give the $e^{-\lambda}$ values for $x = 0, 1, 2, \ldots$ for a given λ and thus facilitate the calculation work.

5. Variance of the Binomial distribution is $n.\ p.\ q.$ and the variance of Poisson distribution is λ. Therefore, $\lambda = n.\ p.\ q.$ Since q is almost equal to unity and as pointed out earlier $n.\ p. = \lambda$ in Poisson distribution. Hence, variance of Poisson distribution is also λ.

NOTES

Example 7.2: Suppose that a manufactured product has 2 defects per unit of product inspected. Use Poisson distribution and calculate the probabilities of finding a product without any defect, with 3 defects and with four defects.

Solution: The product has 2 defects per unit of product inspected. Hence, λ = 2.

Poisson probability function is as,

$$x = 0, 1, 2, \ldots$$

Using the above probability function, we find the required probabilities as,

$$P(\text{without any defects, i.e., } x = 0) = \frac{2^0 . e^{\ \ 2}}{0}$$

$$=$$

$$P(\text{with 3 defects, i.e., } x = 3) = \frac{2^3 . e^{\ \ 2}}{3} =$$

$$=$$

$$P(\text{with 4 defects, i.e., } x = 4) = \frac{2^4 . e^{\ \ 2}}{4} =$$

$$=$$

Example 7.3: How would you use a Poisson distribution to find approximately the probability of exactly 5 successes in 100 trials the probability of success in each trial being $p = 0.1$?

Solution: In the question we have been given,

$$n = 100 \text{ and } p = 0.1$$

$$\therefore \quad \lambda = n.p = 100 \times 0.1 = 10$$

To find the required probability, we can use Poisson probability function as an approximation to Binomial probability function as,

$$f\left(X_i = x\right) = \frac{\lambda^x . e^{-\lambda}}{x!} = \frac{(n.p)^x \ . e^{-(n.p)}}{x!}$$

or,

$$P(5)^7 = \frac{10^5 . e^{\ \ 10}}{5} \quad \frac{100000 \quad 0.00005}{5 \quad 4 \quad 3 \quad 2 \quad 1} \quad \frac{5.00000}{5 \quad 4 \quad 3 \quad 2 \quad 1}$$

$$= \qquad = 0.042$$

2.7 UNIFORM AND NORMAL DISTRIBUTION

Among all the probability distributions, the normal probability distribution is by far the most important and frequently used continuous probability distribution. This is so because this distribution fits well in many types of problems. This distribution is of special significance in inferential statistics since it describes probabilistically the link between a statistic and a parameter (i.e., between the sample results and the population from which

Check Your Progress

6. Write the probability function of binomial distribution.
7. What are the parameters of binomial distribution?
8. What is Poisson distribution?
9. Where and when will you use Poisson distribution?

the sample is drawn). The name of Karl Gauss, 18th century mathematician-astronomer, is associated with this distribution and in honour of his contribution, this distribution is often known as the Gaussian distribution.

The normal distribution can be theoretically derived as the limiting form of many discrete distributions. For instance, if in the binomial expansion of $(p + q)^n$, the value of 'n' is infinity and $p = q = \frac{1}{2}$, then a perfectly smooth symmetrical curve would be obtained. Even if the values of p and q are not equal but if the value of the exponent 'n' happens to be very very large, we get a smooth and symmetrical curve of normal probability. Such curves are called normal probability curves (or at times known as normal curves of error) and such curves represent the normal distributions.[6]

The probability function in case of normal probability distribution[7] is given as,

$$f(x) = \frac{1}{\sigma.\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Where, μ = The mean of the distribution

σ^2 = Variance of the distribution

The normal distribution is thus defined by two parameters, viz., μ and σ^2. This distribution can be represented graphically as shown in Figure 7.5.

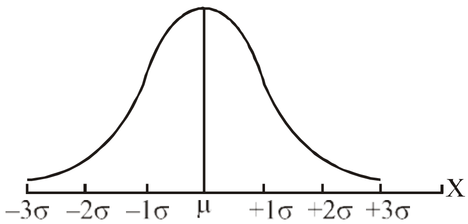


Fig. 7.5 *Curve Representing Normal Distribution*

---

6. Quite often, mathematicians use the normal approximation of the binomial distribution whenever 'n' is equal to or greater than 30 and np and nq each are greater than 5.
7. Equation of the normal curve in its simplest form is,

$$y \quad y_0.e^{x^2/2\ ^2}$$

Where, y = The computed height of an ordinate at a distance of X from the mean.

y_0 = The height of the maximum ordinate at the mean. It is a constant in the equation and is worked out as under:

$$y_0 = \frac{N_i}{\sigma\sqrt{2\pi}}$$

Where, N = Total number of items in the sample

i = Class interval

π = 3.1416

$\therefore$ = $\sqrt{2}$ $\sqrt{6.2832}$ 2.5066

and e = 2.71828, base of natural logarithms

σ = Standard deviation

x = Any given value of the dependent variable expressed as a deviation from the mean.

NOTES

2.7.1 Characteristics of Normal Distribution

The characteristics of the normal distribution or that of normal curve are as follows:

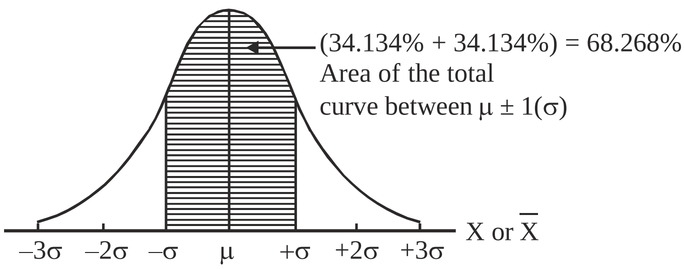
2.7.1.1 It is a symmetric distribution.[8]

2.7.1.2 The mean μ defines where the peak of the curve occurs. In other words, the ordinate at the mean is the highest ordinate. The height of the ordinate at a distance of one standard deviation from mean is 60.653% of the height of the mean ordinate and similarly the height of other ordinates at various standard deviations (σ_s) from mean happens to be a fixed relationship with the height of the mean ordinate.

2.7.1.3 The curve is asymptotic to the base line which means that it continues to approachbut never touches the horizontal axis.

2.7.1.4 The variance (σ^2) defines the spread of the curve.

2.7.1.5 Area enclosed between mean ordinate and an ordinate at a distance of one standard deviation from the mean is always 34.134 per cent of the total area of the curve. It means that the area enclosed between two ordinates at one sigma (S.D.) distance from the mean on either side would always be 68.268 per cent of the total area.This can be shown as follows:



Similarly, the other area relationships are as follows:

Between		*Area Covered to Total Area of the Normal Curve*[9]
μ ± 1	S.D.	68.27%
μ ± 2	S.D.	95.45%
μ ± 3	S.D.	99.73%
μ ± 1.96	S.D.	95%
μ ± 2.578	S.D.	99%
μ ± 0.6745	S.D.	50%

---

8. A symmetric distribution is one which has no skewness. As such it has the following statistical properties:
   (a) Mean=Mode=Median (i.e., $X=Z=M$)
   (b) (Upper Quantile – Median)=(Median – Lower Quantile) (i.e., $Q_3 – M = M – Q_1$)
   (c) Mean Deviation=0.7979(Standard Deviation)
   (d) $\frac{Q_3 \quad Q_1}{2}$ 0.6745 (Standard Deviation)

9. This also means that in a normal distribution, the probability of area lying between variouslimits are as follows:

Limits	Probability of area lying within the stated limits
μ ± 1 S.D.	0.6827
μ ± 2 S.D.	0.9545
μ ± 3 S.D.	0.9973

(This means that almost all cases lie within μ ± 3 S.D. limits)

2.7.1.6 The normal distribution has only one mode since the curve has a single peak. In other words, it is always a unimodal distribution.

2.7.1.7 The maximum ordinate divides the graph of normal curve into two equal parts.

2.7.1.8 In addition to all the above stated characteristics the curve has the following properties:

2.7.1.8.1 $\mu = \bar{x}$

2.7.1.8.2 $\mu_2 = \sigma^2 = \text{Variance}$

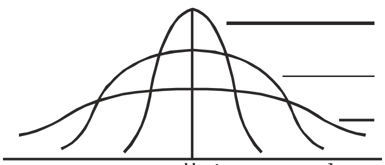
c. $\mu_4 = 3\sigma^4$

d. Moment Coefficient of Kurtosis = 3

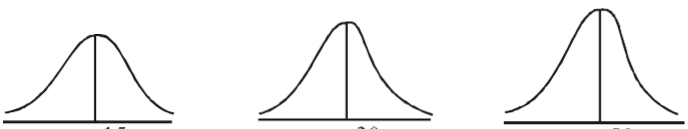
2.7.2 Family of Normal Distributions

We can have several normal probability distributions but each particular normal distribution is being defined by its two parameters viz., the mean (μ) and the standard deviation (s). There is, thus, not a single normal curve but rather a family of normal curves. We can exhibit some of these as follows:

2.7.2.1 Normal curves with identical means but different standard deviations:

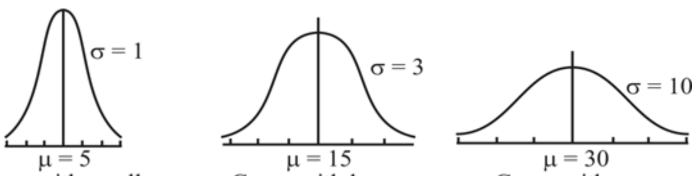


2.7.2.2 Normal curves with identical standard deviation but each with different means:



$\mu = 15$ Curve *A* with smallest mean | $\mu = 30$ Curve *B* with mean between means of curve *A* and curve *C* | $\mu = 50$ Curve *C* with the largest mean

2.7.2.3 Normal curves each with different standard deviations and different means:



$\mu = 5$ Curve with smaller mean and smaller standard deviation | $\mu = 15$ Curve with larger mean and larger standard deviation | $\mu = 30$ Curve with very large mean and very large standard deviation

2.7.3 How to Measure the Area Under the Normal Curve

We have stated above some of the area relationships involving certain intervals of standard deviations (plus and minus) from the means that are true in case of a normal curve. But what should be done in all other cases? We can make use of the statistical tables constructed by mathematicians for the purpose. Using these tables we can find the area (or probability, taking the entire area of the curve as equal to 1) that the normally distributed random variable will lie within certain distances from the mean. These distances are

Self-Instructional

NOTES

defined in terms of standard deviations. While using the tables showing the area under the normal curve we talk in terms of standard variate (symbolically Z) which really means standard deviations without units of measurement and this 'Z' is worked out as,

$$Z = \frac{X - \mu}{\sigma}$$

Where, Z = The standard variate (or number of standard deviations from X to the mean of the distribution)

X = Value of the random variable under consideration

μ = Mean of the distribution of the random variable

σ = Standard deviation of the distribution

The table showing the area under the normal curve (often termed as the standard normal probability distribution table) is organized in terms of standard variate (or Z) values. It gives the values for only half the area under the normal curve, beginning with $Z = 0$ at the mean. Since the normal distribution is perfectly symmetrical the values true for one half of the curve are also true for the other half. We now illustrate the use of such a table for working out certain problems (refer the following examples).

Example 7.4: A banker claims that the life of a regular saving account opened with his bank averages 18 months with a standard deviation of 6.45 months. Answer the following: (a) What is the probability that there will still be money in 22 months in a savings account opened with the said bank by a depositor? (b) What is the probability that the account will have been closed before two years?

Solution:

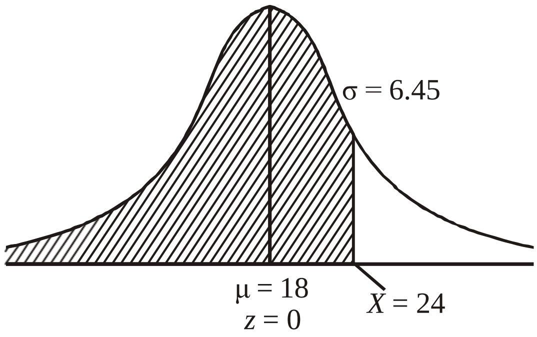
(a) For finding the required probability we are interested in the area of the portion of the normal curve as shaded and shown below:

Let us calculate Z as under:

$$Z = \frac{X - \mu}{\sigma} = \frac{22 - 18}{6.45} = 0.62$$

The value from the table showing the area under the normal curve for $Z = 0.62$ is 0.2324. This means that the area of the curve between $\mu = 18$ and $X = 22$ is 0.2324. Hence, the area of the shaded portion of the curve is (0.5) – (0.2324) = 0.2676 since the area of the entire right hand portion of the curve always happens to be 0.5. Thus the probability that there will still be money in 22 months in a savings account is 0.2676.

(b) For finding the required probability we are interested in the area of the portion of the normal curve as shaded and shown in the following figure:



For the purpose we calculate,

$$Z \quad \frac{24 \quad 18}{6.45} \quad 0.93$$

The value from the concerning table, when $Z = 0.93$, is 0.3238 which refers to the area of the curve between $\mu = 18$ and $X = 24$. The area of the entire left hand portion of the curve is 0.5 as usual.

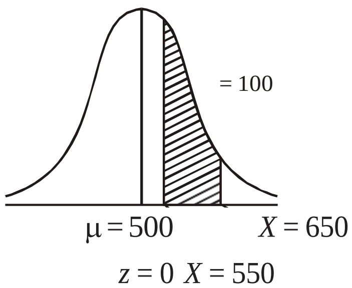
Hence, the area of the shaded portion is (0.5) + (0.3238) = 0.8238 which is the required probability that the account will have been closed before two years, i.e., before 24 months.

Example 7.5: Regarding a certain normal distribution concerning the income of the individuals we are given that mean = 500 rupees and standard deviation =100 rupees. Find the probability that an individual selected at random will belong to income group,

(a) ` 550 – ` 650 (b) ` 420 – 570

Solution:

(a) For finding the required probability we are interested in the area of the portion of the normal curve as shaded and shown below:



For finding the area of the curve between X = 550 – 650, let us do the following calculations:

$$Z = \frac{550 - 500}{100} = \frac{50}{100} = 0.50$$

Corresponding to which the area between $\mu = 500$ and $X = 550$ in the curve as per table is equal to 0.1915 and,

$$Z \quad \frac{650 \quad 500}{100} \quad \frac{150}{100} \quad 1.5$$

Corresponding to which, the area between $\mu = 500$ and $X = 650$ in the curve, as per table, is equal to 0.4332.

Hence, the area of the curve that lies between $X = 550$ and $X = 650$ is,

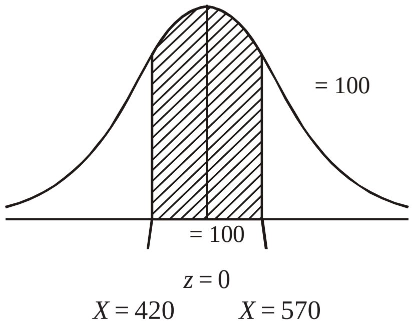
$$(0.4332) - (0.1915) = 0.2417$$

This is the required probability that an individual selected at random will belong to the income group of ` 550 to ` 650.

NOTES

(b) For finding the required probability we are interested in the area of the portion of the normal curve as shaded and shown below:

To find the area of the shaded portion we make the following calculations:



$$Z = \frac{570 - 500}{100} = 0.70$$

Corresponding to which the area between $\mu = 500$ and $X = 570$ in the curve as per table is equal to 0.2580.

and, $$Z = \frac{420 - 500}{100} = -0.80$$

Corresponding to which the area between $\mu = 500$ and $X = 420$ in the curve as per table is equal to 0.2881.

Hence, the required area in the curve between $X = 420$ and $X = 570$ is,

$$(0.2580) + (0.2881) = 0.5461$$

This is the required probability that an individual selected at random will belong to income group of ` 420 to ` 570.

Example 7.6: A certain company manufactures $1\frac{1''}{2}$ all-purpose rope made from imported hemp. The manager of the company knows that the average load-bearing capacity of the rope is 200 lbs. Assuming that normal distribution applies, find the standard deviation of load-bearing capacity for the rope if it is given that the rope has a 0.1210 probability of breaking with 68 lbs. or less pull.

Solution: Given information can be depicted in a normal curve as shown below:

If the probability of the area falling within $\mu = 200$ and $X = 68$ is 0.3790 as stated above, the corresponding value of Z as per the table[10] showing the area of the normal curve is – 1.17 (minus sign indicates that we are in the left portion of the curve)

Now to find σ, we can write,

$$Z = \frac{X - \mu}{\sigma}$$

or, $$-1.17 = \frac{68 - 200}{\sigma}$$

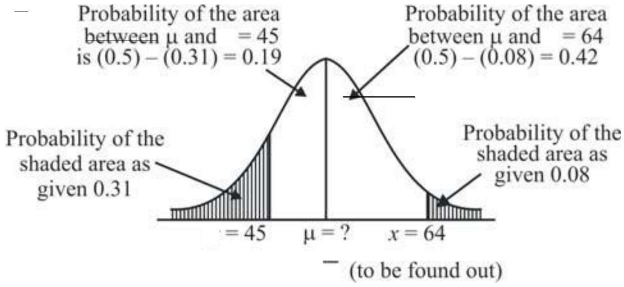
or, $$1.17\sigma \quad 132$$

or, $$\sigma = 112.8 \text{ lbs. approx.}$$

Thus, the required standard deviation is 112.8 lbs. approximately.

Example 7.7: In a normal distribution, 31 per cent items are below 45 and 8 per cent are above 64. Find the $\bar{X}$ and σ of this distribution.

Solution: We can depict the given information in a normal curve as shown below:



If the probability of the area falling within μ and $X = 45$ is 0.19 as stated above, the corresponding value of Z from the table showing the area of the normal curve is – 0.50. Since, we are in the left portion of the curve, we can express this as under,

$$-0.50 = \frac{45 - \mu}{\sigma} \qquad ...(1)$$

Similarly, if the probability of the area falling within μ and $X = 64$ is 0.42, as stated above, the corresponding value of Z from the area table is, +1.41. Since, we are in the right portion of the curve we can express this as under,

$$1.41 = \frac{64 - \mu}{\sigma} \qquad ...(2)$$

If we solve Equations (1) and (2) above to obtain the value of μ or $\bar{X}$, we have,

$$-0.5\,\sigma = 45 - \mu \qquad ...(3)$$

$$1.41\,\sigma = 64 - \mu \qquad ...(4)$$

By subtracting Equation (4) from Equation (3) we have,

$$-1.91\,\sigma = -19$$

$$\therefore \quad \sigma = 10$$

Putting $\sigma = 10$ in Equation (3) we have,

$$-5 = 45 - \mu$$

$$\therefore \quad \mu = 50$$

Hence, $\bar{X}$ (or μ)=50 and σ =10 for the concerning normal distribution.

10. The table is to be read in the reverse order for finding Z value (See Appendix).

2.8 PROBLEMS RELATING TO PRACTICAL APPLICATIONS

NOTES

2.8.1 Fitting a Binomial Distribution

When a binomial distribution is to be fitted to the given data, then the following procedure is adopted:

2.8.1.1 Determine the values of 'p' and 'q' keeping in view that = $n. p.$ and

$q = (1 - p)$.

2.8.1.2 Find the probabilities for all possible values of the given random variable applyingthe binomial probability function, viz.,

$$f(X_i = r) = {}^nC_r p^r q^{n-r}$$

$r = 0, 1, 2, \ldots n$

2.8.1.3 Work out the expected frequencies for all values of random variable bymultiplying N (the total frequency) with the corresponding probability.

2.8.1.4 The expected frequencies, so calculated, constitute the fitted binomial distributionto the given data.

2.8.2 Fitting a Poisson Distribution

When a Poisson distribution is to be fitted to the given data, then the following procedure is adopted:

2.8.2.1 Determine the value of λ, the mean of the distribution.

2.8.2.2 Find the probabilities for all possible values of the given random variable usingthe Poisson probability function, viz.,

$x = 0, 1, 2, \ldots$

2.8.2.3 Work out the expected frequencies as,

$n.p.(X_i = x)$

2.8.2.4 The result of Case (*iii*) above is the fitted Poisson distribution of the given data.

2.8.3 Poisson Distribution as an Approximation of Binomial

Under certain circumstances, Poisson distribution can be considered as a reasonable approximation of

binomial distribution and can be used accordingly. The circumstances which permit this, are when '*n*' is large approaching infinity and *p* is small approaching zero (*n* = number of trials, *p* = probability of 'success'). Statisticians usually take the meaning of large *n*, for this purpose, when $n \geq 20$ and by small '*p*' they mean when $p \leq 0.05$. In cases where these two conditions are fulfilled, we can use mean of the binomial distribution (viz., *n.p.*) in place of the mean of Poisson distribution (viz., λ) so that the probability function of Poisson distribution becomes as stated,

$$f(X_i = x) = \frac{(n.p)^x . e^{-np}}{x!}$$

We can explain Poisson distribution as an approximation of the Binomial distribution with the help of Example 7.8 and Example 7.9.

Example 7.8: The following information is given:

(a) There are 20 machines in a certain factory, i.e., $n = 20$.

(b) The probability of machine going out of order during any day is 0.02.

What is the probability that exactly 3 machines will be out of order on the same day? Calculate the required probability using both Binomial and Poissons Distributions and state whether Poisson distribution is a good approximation of the Binomial distribution in this case.

Solution:

Probability, as per *Poisson probability function* (using *n.p* in place of λ)

(since $n \geq 20$ and $p \leq 0.05$)

$$f\left(X_i = x\right) = \frac{(n.p)^x \, .e^{-np}}{x!}$$

Where, x refers to the number of machines becoming out of order on the same day.

$$P(X_i = 3) = \frac{20 \quad 0.02 \;^3 .e \quad ^{20 \;\; 0.02}}{3}$$

$$= \frac{0.4 \;^3 . \; 0.67032}{3 \quad 2 \quad 1} \quad \frac{(0.064)(0.67032)}{6}$$

$= 0.00715$

Probability, as per *Binomial probability function,*

$$f(X_i = r) = {}^nC_r p^r q^{n-r}$$

Where, $n = 20$, $r = 3$, $p = 0.02$ and hence $q = 0.98$

$$\therefore \quad f(X_i = 3) = {}^{20}C_3 (0.02)^3 \, (0.98)^{17}$$

$= 0.00650$

The difference between the probability of 3 machines becoming out of order on the same day calculated using probability function and binomial probability function is just 0.00065. The difference being very very small, we can state that in the given case Poisson distribution appears to be a good approximation of Binomial distribution.

Example 7.9: How would you use a Poisson distribution to find approximately the probability of exactly 5 successes in 100 trials the probability of success in each trial being $p = 0.1$?

Solution:

In the question we have been given,

$n = 100$ and $p = 0.1$

$$\therefore \quad \lambda = n.p = 100 \times 0.1 = 10$$

To find the required probability, we can use Poisson probability function as an approximation to Binomial probability function, as shown below:

NOTES

or, $P(5)^7 = \frac{10^5.e^{\ 10}}{5} \quad \frac{100000 \quad 0.00005}{5 \quad 4 \quad 3 \quad 2 \quad 1} \quad \frac{5.00000}{5 \quad 4 \quad 3 \quad 2 \quad 1}$

$= \qquad = 0.042$

2.9 BETA DISTRIBUTION

In probability theory and statistics, the **Beta distribution** is a family of continuous probability distributions defined on the interval [0, 1] parameterized by two positive shape parameters, denoted by α and β that appear as exponents of the random variable and control the shape of the distribution.

The beta distribution has been applied to model the behaviour of random variables limited to intervals of finite length in a wide variety of disciplines. For example, it has been used as a statistical description of allele frequencies in population genetics, time allocation in project management or control systems, sunshine data, variability of soil properties, proportions of the minerals in rocks in stratigraphy and heterogeneity in the probability of HIV transmission.

In Bayesian inference, the Beta distribution is the conjugate prior probability distribution for the Bernoulli, Binomial and Geometric distributions. For example, the Beta distribution can be used in Bayesian analysis to describe initial knowledge concerning probability of success, such as the probability that a space vehicle will successfully complete a specified mission. The Beta distribution is a suitable model for the random behavior of percentages and proportions.

The usual formulation of the Beta distribution is also known as the Beta distribution of the first kind, whereas Beta distribution of the second kind is an alternative name for the Beta prime distribution.

The probability density function of the Beta distribution, for $0 < x < 1$, and shape parameters $\alpha, \beta > 0$, is a power function of the variable x and of its reflection $(1-x)$ like follows:

$$= \frac{x^{\alpha-1}(1-x)^{\beta-1}}{\int_0^1 u^{\alpha-1}(1-u)^{\beta-1}\,du}$$

$$= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

$$= \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

Where, $\Gamma(z)$ is the Gamma function. The Beta function, B, appears to be normalization constant to ensure that the total probability integrates to 1. In the above equations x is a realization — an observed value that actually occurred — of a random process X.

The Cumulative Distribution Function (CDF) is given below:

$$F(x;\alpha,\beta) = \frac{B(x;\alpha,\beta)}{B(\alpha,\beta)} = I_x(\alpha,\beta)$$

Where, B (x; α, β) is the incomplete beta function and I_x (α, β) is the regularized incomplete beta function.

The mode of a beta distributed random variable X with α, β > 1 is given by the following expression:

$$\frac{\alpha - 1}{\alpha + \beta - 2}.$$

When both parameters are less than one (α, β < 1), this is the anti-mode - the lowest point of the probability density curve.

The median of the beta distribution is the unique real number for which the regularized incomplete beta function I_x (α, β) = 1/2, there are no general closed form expression for the median of the beta distribution for arbitrary values of α and β. Closed form expressions for particular values of the parameters a and b follow:

- For symmetric cases α = β, median = 1/2.
- For α = 1 and β > 0, median = (this case is the mirror-image of the power function [0, 1] distribution).
- For α > 0 and β = 1, median = (this case is the power function [0, 1] distribution).
- For α = 3 and β = 2, median = 0.6142724318676105..., the real solution to the quartic equation $1–8x^3+6x^4 = 0$, which lies in [0, 1].
- For α = 2 and β = 3, median = 0.38572756813238945... = 1–median (Beta (3, 2)).

The following are the limits with one parameter finite (non zero) and the other approaching these limits:

$$\lim_{\beta \to 0} \text{median} = \lim_{\alpha \to \infty} \text{median} = 1,$$

$$\lim_{\alpha \to 0} \text{median} = \lim_{\beta \to \infty} \text{median} = 0.$$

A reasonable approximation of the value of the median of the Beta distribution, for both α and β greater or equal to one, is given by the following formula:

$$\text{median} \approx \frac{\alpha - \frac{1}{3}}{\alpha + \beta - \frac{2}{3}} \text{ for } \alpha, \beta \geq 1.$$

When α, β ≥ 1, the relative error (the absolute error divided by the median) in this approximation is less than 4% and for both α ≥ 2 and β ≥ 2 it is less than 1%. The absolute error divided by the difference between the mean and the mode is similarly small.

The expected value (mean) (μ) of a beta distribution random variable X with two parameters α and β is a function of only the ratio β/α of these parameters:

$$\mu = E[X]$$

$$= \int_0^1 x \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)} dx$$

$$= \frac{\alpha}{\alpha + \beta}$$

NOTES

Letting $\alpha = \beta$ in the above expression one obtains $\mu = 1/2$, showing that for $\alpha = \beta$ the mean is at the center of the distribution: it is symmetric. Also, the following limits can be obtained from the above expression:

Therefore, for $\beta/\alpha \to 0$, or for $\alpha/\beta \to \infty$, the mean is located at the right end, $x = 1$. For these limit ratios, the Beta distribution becomes a one-point degenerate distribution with a Dirac Delta function spike at the right end, $x = 1$, with probability 1 and zero probability everywhere else. There is 100% probability (absolute certainty) concentrated at the right end, $x = 1$.

Similarly, for $\beta/\alpha \to \infty$, or for $\alpha/\beta \to 0$, the mean is located at the left end, $x = 0$. The Beta distribution becomes a 1 point Degenerate distribution with a Dirac Delta function spike at the left end, $x = 0$, with probability 1 and zero probability everywhere else. There is 100% probability (absolute certainty) concentrated at the left end, $x = 0$. Following are the limits with one parameter finite (non zero) and the other approaching these limits:

$$\lim_{\beta \to 0} \mu = \lim_{\alpha \to \infty} \mu = 1$$

$$\lim_{\alpha \to 0} \mu = \lim_{\beta \to \infty} \mu = 0$$

While for typical unimodal distributions with centrally located modes, inflexion points at both sides of the mode and longer tails with Beta (α, β) such that $\alpha, \beta > 2$ it is known that the sample mean as an estimate of location is not as robust as the sample median, the opposite is the case for uniform or 'U-shaped' Bimodal distributions with beta(α, β) such that $(\alpha, \beta \leq 1)$, with the modes located at the ends of the distribution.

The logarithm of the Geometric Mean (G_X) of a distribution with random variable X is the arithmetic mean of ln (X), or equivalently its expected value:

$$\ln G_x = E\,[\ln X]$$

For a Beta distribution, the expected value integral gives:

$$\mathrm{E}[\ln X]$$

$$= \int_0^1 \ln x \, \frac{x^{\alpha-1}(1-x)^{\beta-1}}{\mathrm{B}(\alpha,\beta)} dx$$

$$= \frac{1}{B(\alpha,\beta)} \int_0^1 \frac{\partial x^{\alpha-1}(1-x)^{\beta-1}}{\partial \alpha} dx$$

$$= \frac{1}{B(\alpha,\beta)} \frac{\partial}{\partial \alpha} \int_0^1 x^{\alpha-1}(1-x)^{\beta-1} dx$$

$$= \frac{1}{B(\alpha,\beta)} \frac{\partial B(\alpha,\beta)}{\partial \alpha}$$

$$= \frac{\partial \ln B(\alpha,\beta)}{\partial \alpha}$$

$$= \frac{\partial \ln \Gamma(\alpha)}{\partial \alpha} - \frac{\partial \ln \Gamma(\alpha+\beta)}{\partial \alpha}$$

$$= \Psi(\alpha) - \Psi(\alpha+\beta)$$

Where, ψ is the Digamma Function.

Therefore, the geometric mean of a Beta distribution with shape parameters α and β is the exponential of the Digamma functions of α and β as follows:

$$G_X = e^{E[\ln X]} = e^{\psi(\alpha)-\psi(\alpha+\beta)}$$

While for a Beta distribution with equal shape parameters $\alpha = \beta$, it follows that Skewness = 0 and Mode = Mean = Median = 1/2, the geometric mean is less than 1/2: $0 < G_X < 1/2$. The reason for this is that the logarithmic transformation strongly weights the values of X close to zero, as ln (X) strongly tends towards negative infinity as X approaches zero, while ln (X) flattens towards zero as $X \to 1$.

Along a line $\alpha = \beta$, the following limits apply:

$$\lim_{\alpha=\beta\to 0} G_X = 0$$

$$\lim_{\alpha=\beta\to \infty} G_X = \frac{1}{2}$$

Following are the limits with one parameter finite (non zero) and the other approaching these limits:

$$\lim_{\beta\to 0} G_X = \lim_{\alpha\to\infty} G_X$$

$$\lim_{\alpha\to 0} G_X = \lim_{\beta\to\infty} G_X = 0$$

The accompanying plot shows the difference between the mean and the geometric mean for shape parameters α and β from zero to 2. Besides the fact that the difference between them approaches zero as α and β approach infinity and that the difference becomes large for values of α and β approaching zero, one can observe an evident asymmetry of the geometric mean with respect to the shape parameters α and β. The difference between the geometric mean and the mean is larger for small values of α in relation to β than when exchanging the magnitudes of β and α.

The inverse of the Harmonic Mean (H_X) of a distribution with random variable X is the arithmetic mean of $1/X$, or, equivalently, its expected value. Therefore, the Harmonic Mean (H_X) of a beta distribution with shape parameters α and β is:

NOTES

NOTES

$$= \frac{1}{\int_0^1 \frac{x^{\alpha-1}(1-x)^{\beta-1}}{x\mathrm{B}(\alpha,\beta)}dx}$$

$$= \frac{\alpha-1}{\alpha+\beta-1} \text{ if } \alpha > 1 \text{ and } \beta > 0$$

The Harmonic Mean (H_X) of a beta distribution with $\alpha < 1$ is undefined, because its defining expression is not bounded in [0, 1] for shape parameter a less than unity.

Letting $\alpha = \beta$ in the above expression one can obtain the following:

$$H_X = \frac{\alpha-1}{2\alpha-1},$$

Showing that for $\alpha = \beta$ the harmonic mean ranges from 0, for $\alpha = \beta = 1$, to 1/2, for $\alpha = \beta \to \infty$.

Following are the limits with one parameter finite (non zero) and the other approaching these limits:

The Harmonic mean plays a role in maximum likelihood estimation for the four parameter case, in addition to the geometric mean. Actually, when performing maximum likelihood estimation for the four parameter case, besides the harmonic mean H_X based on the random variable X, also another harmonic mean appears naturally: the harmonic mean based on the linear transformation (1"X), the mirror image of X, denoted by $H_{(1-X)}$:

The Harmonic mean ($H_{(1-X)}$) of a Beta distribution with $\beta < 1$ is undefined, because its defining expression is not bounded in [0, 1] for shape parameter β less than unity.

Using $\alpha = \beta$ in the above expression one can obtain the following:

$$H_{(1-X)} = \frac{\beta-1}{2\beta-1},$$

This shows that for $\alpha = \beta$ the harmonic mean ranges from 0, for $\alpha = \beta = 1$, to 1/2, for $\alpha = \beta \to \infty$.

Following are the limits with one parameter finite (non zero) and the other approaching these limits:

Although both H_X and $H_{(1-X)}$ are asymmetric, in the case that both shape parameters are equal $\alpha = \beta$, the harmonic means are equal: $H_X = H_{(1-X)}$. This equality follows from the following symmetry displayed between both harmonic means:

NOTES

2.10 GAMMA DISTRIBUTION

In probability theory and statistics, the gamma distribution is defined as a two parameter family of continuous probability distributions. The special cases of the gamma distribution include common exponential distribution and Chi-squared distribution. The following are three significant parametrizations of gamma distribution that are commonly used:

1. With a shape parameter k and a scale parameter θ.
2. With a shape parameter $= k$ and an inverse scale parameter $\beta = 1/\theta$, called a rate parameter.
3. With a shape parameter k and a mean parameter $\mu = k/\beta$.

In each of the above mentioned three forms, both parameters are positive real numbers.

$H_X(B(\alpha,\beta)) = H_{(1-X)}(B(\beta,\alpha))$ if $\alpha,\beta > 1$.

$\lim_{\beta\to 1} H_{(1-X)} =$ undefined, $\lim_{\alpha\to\infty} H_{(1-X)} = 0$, $\lim_{\alpha\to 0} H_{(1-X)} = \lim_{\beta\to\infty} H_{(1-X)} = 1$

The parameterization with k and θ is basically used in econometrics and some specific applied fields where the gamma distribution is normally used to model waiting times. The parameterization with α and β is basically used in Bayesian statistics, where the gamma distribution is typically used as a conjugate prior distribution for several types of inverse scaling or rate parameters, such as using the λ of an exponential distribution or a Poisson distribution or the β of the gamma distribution itself. If k is an integer, then the distribution represents an Erlang distribution as the sum of k independent exponentially distributed random variables, each of which has a mean of θ.

Some Significant Properties of Gamma Distribution

The following are some significant properties of gamma distribution:

1. A random variable X that is gamma distributed with shape k and scale θ is denoted by:

$$X \sim \Gamma(k, \theta) \equiv \text{Gamma}(k, \theta)$$

2. The probability density function using the shape scale parametrization is denoted by:

$$f(x; k, \theta) = \qquad \text{for } x > 0 \text{ and } k, \theta > 0.$$

Here $\Gamma(k)$ is the gamma function evaluated at k.

3. If k is a positive integer, i.e., the distribution is an Erlang distribution then it is defined as follows:

$$F(x;\ k,\ \theta) =$$

NOTES

4. The gamma distribution can be parameterized in terms of a shape parameter $\alpha = k$ and an inverse scale parameter $\beta = 1/\theta$, called a rate parameter. A random variable X that is gamma distributed with shape α and rate β is denoted by:

$$X \sim \Gamma(\alpha, \beta) \equiv \text{Gamma}(\alpha, \beta)$$

5. The corresponding probability density function in the shape rate parametrization is denoted by:

$$g(x;\ \alpha,\ \beta) = \frac{x \quad ^{1} e \quad ^{x}}{\qquad} \quad \text{for } x \geq 0 \text{ and } \alpha, \beta > 0$$

6. The function f below is a probability density function for any k>0.

$$f(x) = \qquad\qquad 0 < x < \infty$$

7. A random variable X with this probability density function is said to have the *gamma distribution* with *shape parameter k.*
8. The gamma probability density function satisfies the following properties:
    - If $0<k<1$ then f is decreasing with $f(x)\to\infty$ as $x\downarrow 0$.
    - If k=1 then f is decreasing with $f(0)=1$.
    - If $k>1$ then f increases on the interval $(0, k–1)$ and decreases on the interval $(k–1, \infty)$.
    - The special case k=1 gives the *standard exponential distribution*. When $k\geq 1$, then the distribution is unimodal with mode $k–1$.

2.11 SUMMARY

- The expected value (or mean) of X is the weighted average of the possible values that X can take.
- The variance of a random variable tells us something about the spread of the possible values of the variable. For a discrete random variable X, the variance of X is written as Var(X).
- Mean of random variable is the sum of the values of the random variable weighted by the probability that the random variable will take on the value.
- The mean or the expected value of random variable may not be adequate enough at times to study the problem as to how random variable actually behaves and we may as well be interested in knowing something about how the values of random variable are dispersed about the mean.
- The standard distribution is a special case of the normal distribution. It is the distribution that occurs when a normal random variable has a mean of zero and a standard deviation of one.

Check Your Progress

10. Why a normal distribution is preferred over the other types?
11. Under what circumstances, Poisson distribution is considered as an approximation of binomial distribution?

NOTES

- Statistical inference means drawing conclusions based on data that is subjected to random variation; for example, sampling variation or observational errors.
- Binomial distribution (or the Binomial probability distribution) is a widely used probability distribution concerned with a discrete random variable and as such is an example of a discrete probability distribution.
- Poisson distribution is frequently used in context of Operations Research and for this reason has a great significance for management people.
- Unlike binomial distribution, Poisson distribution cannot be deducted on purely theoretical grounds based on the conditions of the experiment.
- Among all the probability distributions, the normal probability distribution is by far the most important and frequently used continuous probability distribution. This is so because this distribution fits well in many types of problems.
- Under certain circumstances, Poisson distribution can be considered as a reasonable approximation of binomial distribution and can be used accordingly. The circumstances which permit this, are when '*n*' is large approaching infinity and *p* is small approaching zero (*n* = Number of trials, *p* = Probability of 'success').

2.12 KEY TERMS

- **Expected value of random variable:** The average value that would occur if we have to average an infinite number of outcomes of the random variable
- **Standard distribution:** The distribution that occurs when a normal random variable has a mean of zero and a standard deviation of one
- **Statistical model:** A set of assumptions concerning the generation of the observed data and similar data
- **Binomial distribution:** Also called as Bernoulli process and is used to describe discrete random variable
- **Poisson distribution:** Used to describe the empirical results of past experiments relating to the problem and plays an important role in queuing theory, inventory control problems and risk models

2.13 ANSWERS TO 'CHECK YOUR PROGRESS'

1. The expected value is the sum of each of the possible outcomes and the probability of the outcome occurring.
2. Mean of random variable is the sum of the values of the random variable weighted by the probability that the random variable will take on the value.
3. The expected value of a random variable is calculated by weighting each value of a random variable by its probability and summing over all values.
4. The normal random variable of a standard normal distribution is called a standard score or a *z*-score.
5. Statistical inference means drawing conclusions based on data that is subjected to random variation; for example, sampling variation or observational errors.

NOTES

6. The probability function of the binomial distribution is written as,

$$f(X = r) = {}^nC_r\, p^r q^{n-r}$$
$$r = 0, 1, 2\ldots n$$

Where, n = Numbers of trials

p = Probability of success in a single trial.

$q = (1 - p)$ = Probability of 'failure' in a single trial.

r = Number of successes in 'n' trials.

7. The parameters of binomial distribution are p and n, where p specifies the probability of success in a single trial and n specifies the number of trials.
8. Poisson distribution is a discrete probability distribution that is frequently used in the context of Operations Research. Unlike binomial distribution, Poisson distribution cannot be deduced on purely theoretical grounds based on the conditions of the experiment. In fact, it must be based on experience, i.e., on the empirical results of past experiments relating to the problem under study.
9. Poisson distribution is used when the probability of happening of an event is very small and n is very large such that the average of series is a finite number. This distribution is good for calculating the probabilities associated with X occurrences in a given time period or specified area.
10. Normal distribution is the most important and frequently used continuous probability distribution among all the probability distributions. This is so because this distribution fits well in many types of problems. This distribution is of special significance in inferential statistics since it describes probabilistically the link between a statistic and a parameter.
11. When n is large approaching infinity and p is small approaching zero, Poisson distribution is considered as an approximation of binomial distribution.

2.14 QUESTIONS AND EXERCISES

Short-Answer Questions

1. Differentiate between conditional and iterated expectations.
2. Differentiate between a discrete and a continuous variable.
3. How does a random variable function?
4. What do you understand by variance of random variable?
5. What do you understand by the term standard distribution?
6. Differentiate between statistical inference and descriptive statistics.
7. What are the characteristics of Bernoulli process?
8. What is Poisson distribution? What are its important measures?
9. When is the Poisson distribution used?
10. Define the term normal distribution. What are the characteristics of normal distribution?
11. Write the formula for measuring the area under the curve.
12. Under what circumstances the normal probability distribution can be used?

Long-Answer Questions

NOTES

1. Explain the concept of expectation of a random variable with the help of an example.
2. *A* and *B* roll a die. Whoever gets a 6 first, wins ` 550. Find their individual expectations if *A* makes a start. What will be the answer if *B* makes a start?
3. What is statistical inference? Discuss in detail.
4. Describe binomial distribution and its measures.
5. The following is a probability distribution:

X_i	$pr(X_i)$
0	1/8
1	2/8
2	3/8
3	2/8

Calculate the expected value of X_i, its variance, and standard deviation.

6. A coin is tossed 3 times. Let *X* be the number of runs in the sequence of outcomes: first toss, second toss, third toss. Find the probability distribution of *X*. What values of *X* are most probable?
7. (a) Explain the meaning of Bernoulli process pointing out its main characteristics.
   (b) Give a few examples narrating some situations wherein binomial *pr*. distribution can be used.
8. Poisson distribution can be an approximation of binomial distribution. Explain.
9. State the distinctive features of the binomial, Poisson and normal probability distributions. When does a binomial distribution tend to become a normal and a Poisson distribution? Explain.
10. Explain the circumstances when the following probability distributions are used:
    (a) Binomial distribution
    (b) Poisson distribution
    (c) Normal distribution
11. Certain articles were produced of which 0.5 per cent are defective, are packed in cartons, each containing 130 articles. What proportion of cartons are free from defective articles? What proportion of cartons contain 2 or more defective?
    (Given $e^{-0.5} = 0.6065$).
12. The following mistakes per page were observed in a book:

No. of Mistakes Per Page	*No. of Times the Mistake Occurred*
0	211
1	90
2	19
3	5
4	0
Total	325

Fit a Poisson distribution to the data given above and test the goodness of fit.

NOTES

13. In a distribution exactly normal, 7 per cent of the items are under 35 and 89 per cent are under 63. What are the mean and standard deviation of the distribution?
14. Assume the mean height of soldiers to be 68.22 inches with a variance of 10.8 inches. How many soldiers in a regiment of 1000 would you expect to be over six feet tall?
15. Fit a normal distribution to the following data:

Height in Inches	*Frequency*
60–62	5
63–65	18
66–68	42
69–71	27
72–74	8

2.15 FURTHER READING

Allen, R.G.D. 2008. *Mathematical Analysis For Economists*. London: Macmillan and Co., Limited.

Chiang, Alpha C. and Kevin Wainwright. 2005. *Fundamental Methods of Mathematical Economics*, 4 edition. New York: McGraw-Hill Higher Education.

Yamane, Taro. 2012. *Mathematics For Economists: An Elementary Survey*. USA: Literary Licensing.

Baumol, William J. 1977. *Economic Theory and Operations Analysis*, 4th revised edition. New Jersey: Prentice Hall.

Hadley, G. 1961. *Linear Algebra*, 1st edition. Boston: Addison Wesley.

Vatsa , B.S. and Suchi Vatsa. 2012. *Theory of Matrices*, 3rd edition. London: New Academic Science Ltd.

Madnani, B C and G M Mehta. 2007. *Mathematics for Economists*. New Delhi: Sultan Chand & Sons.

Henderson, R E and J M Quandt. 1958. *Microeconomic Theory: A Mathematical Approach*. New York: McGraw-Hill.

Nagar, A.L. and R.K.Das. 1997. *Basic Statistics*, 2nd edition. United Kingdom: Oxford University Press.

Gupta, S.C. 2014. *Fundamentals of Mathematical Statistics*. New Delhi: Sultan Chand & Sons.

M. K. Gupta, A. M. Gun and B. Dasgupta. 2008. *Fundamentals of Statistics*. West Bengal: World Press Pvt. Ltd.

Saxena, H.C. and J. N. Kapur. 1960. *Mathematical Statistics*, 1st edition. New Delhi: S. Chand Publishing.

Hogg, Robert V., Joeseph McKean and Allen T Craig. *Introduction to Mathematical Statistics*, 7th edition. New Jersey: Pearson.

UNIT III STATISTICAL INFERENCE

NOTES

Structure

3.0 INTRODUCTION

In this unit, you will learn about the basic concepts of statistical inference. Statistical inference is the process of deducing properties of an underlying distribution by analysis of data. Inferential statistical analysis infers properties about a population: this includes testing hypotheses and deriving estimates. You will also learn about the hypothesis formulation and test of significance. This unit will also discuss the Chi-square statistics, *t*-test and *F*-statistic. Finally, you will learn about the one-tailed and two-tailed tests.

3.1 UNIT OBJECTIVES

After going through this unit, you will be able to:

- Discuss about sampling distribution
- Understand hypothesis distribution and test of significance
- Explain the significance of Chi-square statistics
- Discuss about t-statistics and F-statistics
- Explain the significance of one-tailed and two-tailed tests

3.2 SAMPLING DISTRIBUTION

One of the major objectives of statistical analysis is to know the 'true' values of different parameters of the population. Since it is not possible due to time, cost and other constraints to take the entire population for consideration, random samples are taken from the

NOTES

population. These samples are analysed properly and they lead to generalizations that are valid for the entire population. The process of relating the sample results to population is referred to as, 'Statistical Inference' or 'Inferential Statistics'.

In general, a single sample is taken and its mean X is considered to represent the population mean. However, in order to use the sample mean to estimate the population mean, we should examine every possible sample (and its mean, etc.) that could have occurred, because a single sample may not be representative enough. If it was possible to take all the possible samples of the same size, then the distribution of the results of these samples would be referred to as, 'sampling distribution'. The distribution of the means of these samples would be referred to as, 'sampling distribution; of the means'.

The relationship between the sample means and the population mean can best be illustrated by Example 8.1.

Example 8.1: Suppose a babysitter has 5 children under her supervision with average age of 6 years. However, individually, the age of each child be as follows:

$$X_1 = 2$$
$$X_2 = 4$$
$$X_3 = 6$$
$$X_4 = 8$$
$$X_5 = 10$$

Now these 5 children would constitute our entire population, so that N = 5.

Solution:

The population mean $\mu \quad \dfrac{X}{N}$

$$= \frac{2+4+6+8+10}{5} = 30/5 = 6$$

and the standard deviation is given by the formula:

Now, let us calculate the standard deviation.

X	μ	$(X-\mu)^2$
2	6	16
4	6	4
6	6	0
8	6	4
10	6	16
		Total = $\Sigma (X-\mu)^2 = 40$

Then,

$$\sigma \quad \sqrt{\frac{40}{5}} \quad \sqrt{8} \quad 2.83$$

Now, let us assume the sample size, $n = 2$, and take all the possible samples of size 2, from this population. There are 10 such possible samples. These are as follows, along with their means.

X_1, X_2	(2, 4)	$\bar{X}_1$	= 3
X_1, X_3	(2, 6)		= 4
X_1, X_4	(2, 8)		= 5
X_1, X_5	(2,10)		= 6
X_2, X_3	(4, 6)		= 5
X_2, X_4	(4, 8)		= 6
X_2, X_5	(4, 10)		= 7
X_3, X_4	(6, 8)		= 7
X_3, X_5	(6, 10)		= 8
X_4, X_5	(8, 10)		= 9

Now, if only the first sample was taken, the average of the sample would be 3. Similarly, the average of the last sample would be 9. Both of these samples are totally unrepresentative of the population. However, if a grand mean X of the distribution of these sample means is taken, then,

$$\bar{\bar{X}} = \frac{\sum_{i=1}^{10} \bar{X}_i}{10} = \frac{3+4+5+6+5+6+7+7+8+9}{10} = 60/10 = 6$$

This grand mean has the same value as the mean of the population. Let us organize this distribution of sample means into a frequency distribution and probability distribution.

Sample Mean	*Freq.*	*Rel. Freq.*	*Prob.*
3	1	1/10	0.1
4	1	1/10	0.1
5	2	2/10	0.2
6	2	2/10	0.2
7	2	2/10	0.2
8	1	1/10	0.1
9	1	1/10	0.1
			1.00

This probability distribution of the sample means is referred to as 'sampling distribution of the mean.'

Sampling Distribution of the Mean

The sampling distribution of the mean can thus be defined as, 'A probability distribution of all possible sample means of a given size, selected from a population'.

NOTES

Accordingly, the sampling distribution of the means of the ages of children as tabulated in Example 8.1, has 3 predictable patterns. These are as follows:

(i) The mean of the sampling distribution and the mean of the population are equal. This can be shown as follows:

Sample mean ( )	*Prob. P( )*
3	0.1
4	0.1
5	0.2
6	0.2
7	0.2
8	0.1
9	0.1
	1.00

Then,

μ = (3 × 0.l) + (4 × 0.l) + (5 × 0.2) + (6 × 0.2) + (7 × 0.2) + (8 × 0.l) + 9 ×.l) = 6

This value is the same as the mean of the original population.

(ii) The spread of the sample means in the distribution is smaller than in the population values. For example, the spread in the distribution of sample means above is from 3 to 9, while the spread in the population was from 2 to 10.

(iii) The shape of the sampling distribution of the means tends to be, 'Bell- shaped' and approximates the normal probability distribution, even when the population is not normally distributed. This last property leads us to the 'Central Limit Theorem'.

3.2.1 Central Limit Theorem

Central Limit Theorem states that, 'Regardless of the shape of the population, the distribution of the sample means approaches the normal probability distribution as the sample size increases.'

The question now is how large should the sample size be in order for the distribution of sample means to approximate the normal distribution for any type of population. In practice, the sample sizes of 30 or larger are considered adequate for this purpose. This should be noted however, that the sampling distribution would be normally distributed, if the original population is normally distributed, no matter what the sample size.

As we can see from our sampling distribution of the means, the grand mean $\bar{\bar{X}}$ of the sample means or μ equals μ, the population mean. However, realistically speaking, it is not possible to take all the possible samples of size *n* from the population. In practice only one sample is taken, but the discussion on the sampling distribution is concerned with the proximity of 'a' sample mean to the population mean.

It can be seen that the possible values of sample means tend towards the population mean, and according to Central Limit Theorem, the distribution of sample means tend to be normal for a sample size of *n* being larger than 30. Hence, we can draw conclusions based upon our knowledge about the characteristics of the normal distribution.

NOTES

For example, in the case of sampling distribution of the means, if we know the grand mean μ of this distribution, which is equal to p, and the standard deviation of this distribution, known as 'Standard error of free mean' and denoted by σ, then we know from the normal distribution that there is a 68.26 per cent chance that a sample selected at random from a population, will have a mean that lies within one standard error of the mean of the population mean. Similarly, this chance increases to 95.44 per cent, that the sample mean will lie within two standard errors of the mean (σ) of the population mean. Hence, knowing the properties of the sampling distribution tells us as to how close the sample mean will be to the true population mean.

3.2.2 Standard Error

Standard Error of the Mean (σ)

Standard error of the mean (σ) is a measure of dispersion of the distribution of sample means and is similar to the standard deviation in a frequency distribution and it measures the likely deviation of a sample mean from the grand mean of the sampling distribution.

If all sample means are given, then (σ) can be calculated as follows:

where N = Number of sample means

Thus we can calculate $\sigma_{\overline{x}}$ for Example 8.1 of the sampling distribution of the ages of 5 children as follows:

3	6	9
4	6	4
5	6	1
6	6	0
7	6	1
8	6	4
9	6	9
		$\Sigma\ (\overline{x} - \mu_{\overline{x}})^2 = 28$

Then,

$$\sigma_{\overline{x}} = \sqrt{\frac{\Sigma(\overline{x} - \mu_{\overline{x}})}{N}}$$

$$= \sqrt{\frac{28}{7}}$$

However, since it is not possible to take all possible samples from the population, we must use alternate methods to compute σ .

The standard error of the mean can be computed from the following formula, if the population is finite and we know the population mean. Hence, Where,

σ = Population standard deviation
N = Population size
n = Sample size

NOTES

This formula can be made simpler to use by the fact that we generally deal with very large populations, which can be considered infinite, so that if the population size A' is very large and sample size *n* is small, as for example in the case of items tested from assembly line operations, then,

would approach 1.

Hence,

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

The factor $\sqrt{\frac{(N-n)}{(N-n)}}$ is also known as the 'finite correction factor', and should be used when the population size is finite.

As this formula suggests, $\sigma_{\bar{x}}$ decreases as the sample size (w) increases, meaning that the general dispersion among the sample means decreases, meaning further that any single sample mean will become closer to the population mean, as the value of ($\sigma_{\bar{x}}$) decreases. Additionally, since according to the property of the normal curve, there is a 68.26 per cent chance of the population mean being within one $\sigma_{\bar{x}}$ of the sample mean, a smaller value of $\sigma_{\bar{x}}$ will make this range shorter; thus making the population mean closer to the sample mean (refer Example 8.2).

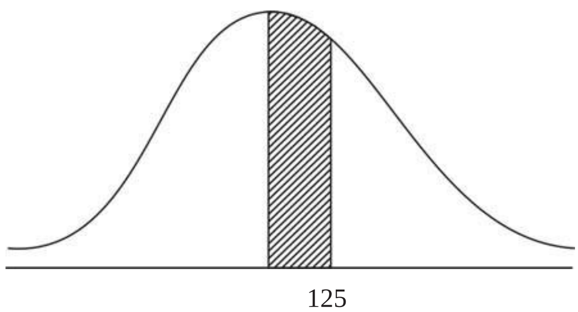
Example 8.2: The *IQ* scores of college students are normally distributed with the mean of 120 and standard deviation of 10.

3.2.2.1 What is the probability that the *IQ* score of any one student chosen at random isbetween 120 and 125?

3.2.2.2 If a random sample of 25 students is taken, what is the probability that the meanof this sample will be between 120 and 125.

Solution:

(a) Using the standardized normal distribution formula,



μ = 120

σ = 10

$$Z = \frac{125 - 120}{10} = 5/10 = .5$$

The area for Z = .5 is 19.15.

NOTES

This means that there is a 19.15 per cent chance that a student picked up at random will have an IQ score between 120 and 125.

(b) With the sample of 25 students, it is expected that the sample mean will be much closer to the population mean, hence it is highly likely that the sample mean would be between 120 and 125.

The formula to be used in the case of standardized normal distribution for sampling distribution of the means is given by,

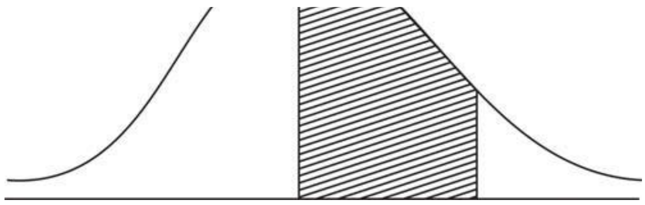
$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$

where,

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{10}{\sqrt{25}} = \frac{10}{5} = 2$$

Hence,

$$Z = \frac{125 - 120}{2} = 5/2 = 2.5$$



$$Z = \frac{\bar{X} - \mu}{\sigma_{\bar{x}}}$$

where,

Then,

The area for Z = 2.5 is 49.38.

This shows that there is a chance of 49.38 per cent that the sample mean will be between 120 and 125. As the sample size increases further, this chance will also increase. It can be noted that the probability of a sample mean being between 120 and 125 is much higher than the probability of an individual student having an IQ between 120 and 125.

3.3 HYPOTHESIS FORMULATION AND TEST OF SIGNIFICANCE

NOTES

A hypothesis is an approximate assumption that a researcher wants to test for its logical or empirical consequences. Hypothesis refers to a provisional idea whose merit needs evaluation, but has no specific meaning, though it is often referred as a convenient mathematical approach for simplifying cumbersome calculation. Setting up and testing hypothesis is an integral art of statistical inference. Hypotheses are often statements about population parameters like variance and expected value. During the course of hypothesis testing, some inference about population like the mean and proportion are made. Any useful hypothesis will enable predictions by reasoning including deductive reasoning. According to Karl Popper, a hypothesis must be falsifiable and that a proposition or theory cannot be called scientific if it does not admit the possibility of being shown false. Hypothesis might predict the outcome of an experiment in a lab, setting the observation of a phenomenon in nature. Thus, hypothesis is an explanation of a phenomenon proposal suggesting a possible correlation between multiple phenomena.

The characteristics of hypothesis are as follows:

- **Clear and Accurate:** Hypothesis should be clear and accurate so as to draw a consistent conclusion.
- **Statement of Relationship between Variables:** If a hypothesis is relational, it should state the relationship between different variables.
- **Testability:** A hypothesis should be open to testing so that other deductions can be made from it and can be confirmed or disproved by observation. The researcher should do some prior study to make the hypothesis a testable one.
- **Specific with Limited Scope:** A hypothesis, which is specific, with limited scope, is easily testable than a hypothesis with limitless scope. Therefore, a researcher should pay more time to do research on such kind of hypothesis.
- **Simplicity:** A hypothesis should be stated in the most simple and clear terms to make it understandable.
- **Consistency:** A hypothesis should be reliable and consistent with established and known facts.
- **Time Limit:** A hypothesis should be capable of being tested within a reasonable time. In other words, it can be said that the excellence of a hypothesis is judged by the time taken to collect the data needed for the test.
- **Empirical Reference:** A hypothesis should explain or support all the sufficient facts needed to understand what the problem is all about.

A hypothesis is a statement or assumption concerning a population. For the purpose of decision-making, a hypothesis has to be verified and then accepted or rejected. This is done with the help of observations. We test a sample and make a decision on the basis of the result obtained. Decision-making plays a significant role in different areas, such as marketing, industry and management.

Statistical Decision-Making

Testing a statistical hypothesis on the basis of a sample enables us to decide whether the hypothesis should be accepted or rejected. The sample data enables us to accept or reject the hypothesis. Since the sample data gives incomplete information about the population, the result of the test need not be considered to be final or unchallengeable. The procedure, on the basis of which sample results, enables to decide whether a hypothesis is to be accepted or rejected. This is called Hypothesis Testing or Test of Significance.

Note 1: A test provides evidence, if any, against a hypothesis, usually called a null hypothesis. The test cannot prove the hypothesis to be correct. It can give some evidence against it. The test of hypothesis is a procedure to decide whether to accept or reject a hypothesis.

Note 2: The acceptance of a hypotheses implies, if there is no evidence from the sample that we should believe otherwise.

The rejection of a hypothesis leads us to conclude that it is false. This way of putting the problem is convenient because of the uncertainty inherent in the problem. In view of this, we must always briefly state a hypothesis that we hope to reject.

A hypothesis stated in the hope of being rejected is called a *null hypothesis* and is denoted by H_0.

If H_0 is rejected, it may lead to the acceptance of an alternative hypothesis denoted by H_1.

For example, a new fragrance soap is introduced in the market. The null hypothesis H_0, which may be rejected, is that the new soap is not better than the existing soap.

Similarly, a dice is suspected to be rolled. Roll the dice a number of times to test.

The null hypothesis H_0: $p = 1/6$ for showing six.

The alternative hypothesis H_1: $p \neq 1/6$.

For example, skulls found at an ancient site may all belong to race *X* or race *Y* on the basis of their diameters. We may test the hypothesis, that the mean is μ of the population from which the present skulls came. We have the hypotheses.

$$H_0 : \mu = \mu_x, H_1 : \mu = \mu_y$$

Here, we should not insist on calling either hypothesis null and the other alternative since the reverse could also be true.

Committing Errors: Type I and Type II

Types of Errors

There are two types of errors in statistical hypothesis, which are as follows:

- **Type I Error:** In this type of error, you may reject a null hypothesis when it is true. It means rejection of a hypothesis, which should have been accepted. It is denoted by α (alpha) and is also known alpha error.
- **Type II Error:** In this type of error, you are supposed to accept a null hypothesis when it is not true. It means accepting a hypothesis, which should have been rejected. It is denoted by β (beta) and is also known as beta error.

  Type I error can be controlled by fixing it at a lower level. For example, if you fix it at 2 per cent, then the maximum probability to commit Type I error is 0.02. However, reducing Type I error has a disadvantage when the sample size is fixed, as it increases the chances of Type II error. In other words, it can be said that both types of errors cannot be reduced simultaneously. The only solution of this problem is to set an appropriate level by considering the costs and penalties attached to them or to strike a proper balance between both types of errors.

NOTES

In a hypothesis test, a Type I error occurs when the null hypothesis is rejected when it is in fact true; that is, H_0 is wrongly rejected. For example, in a clinical trial of a new drug, the null hypothesis might be that the new drug is no better, on average, than the current drug; that is H_0: there is no difference between the two drugs on average. A Type I error would occur if we concluded that the two drugs produced different effects, when in fact there was no difference between them.

In a hypothesis test, a Type II error occurs when the null hypothesis H_0 is not rejected, when it is in fact false. For example, in a clinical trial of a new drug, the null hypothesis might be that the new drug is no better, on average, than the current drug; that is H_0: there is no difference between the two drugs on average. A Type II error would occur if it were concluded that the two drugs produced the same effect, that is, there is no difference between the two drugs on average, when in fact they produced different ones.

In how many ways can we commit errors?

We reject a hypothesis when it may be true. This is Type I Error.

We accept a hypothesis when it may be false. This is Type II Error.

The other true situations are desirable: We accept a hypothesis when it is true. We reject a hypothesis when it is false.

	Accept H_0	**Reject H_0**
H_0 True	Accept True H_0 Desirable	Reject True H_0 Type I Error
H_1 False	Accept False H_0 Type II Error	Reject False H_0 Desirable

The level of significance implies the probability of Type I error. A 5 per cent level implies that the probability of committing a Type I error is 0.05. A 1 per cent level implies 0.01 probability of committing Type I error.

Lowering the significance level and hence the probability of Type I error is good but unfortunately, it would lead to the undesirable situation of committing Type II error.

To Sum Up:

3.3.1 **Type I Error**: Rejecting H_0 when H_0 is true.

3.3.2 **Type II Error**: Accepting H_0 when H_0 is false.

Note: The probability of making a Type I error is the level of significance of a statistical test. It is denoted by α.

Where, α = Prob. (Rejecting H_0 / H_0 true) 1–

α = Prob. (Accepting H_0 / H_0 true)

The probability of making a Type II error is denoted by β.

Where, β = Prob. (Accepting H_0 / H_0 false)

1– = Prob. (Rejecting H_0 / H_0 false) = Prob. (The test correctly rejects H_0 when H_0 is false)

1–β is called the power of the test. It depends on the level of significance α, sample size n and the parameter value.

8.3.1 Test of Significance

NOTES

Tests for a Sample Mean $\bar{X}$

We have to test the null hypothesis that the population mean has a specified value μ, i.e., H_0: $\bar{X}$ = μ. For large n, if H_0 is true then,

$z \quad \left|\dfrac{\bar{X}}{SE(\bar{X})}\right|$ is approximately nominal. The theoretical region for z depending on the desired level of significance can be calculated.

For example, a factory produces items, each weighing 5 kg with variance 4. Can a random sample of size 900 with mean weight 4.45 kg be justified as having been taken from this factory?

$$n = 900$$
$$\bar{X} = 4.45$$
$$\mu = 5$$
$$\sigma = \sqrt{4} = 2$$
$$z \quad \left|\frac{\bar{X}}{SE(\bar{X})}\right| = \left|\frac{\bar{X}}{/\sqrt{n}}\right| = \left|\frac{4.45 \quad 5}{2/30}\right| = 8.25$$

We have $z > 3$. The null hypothesis is rejected. The sample may not be regarded as originally from the factory at 0.27 per cent level of significance (corresponding to 99.73 per cent acceptance region).

Test for Equality of Two Proportions

If P_1, P_2 are proportions of some characteristic of two samples of sizes n_1, n_2, drawn from populations with proportions P_1, P_2, then we have H_0: $P_1 = P_2$ vs H_1:$P_1 \neq P_2$.

- **Case (*I*):** If H_0 is true, then let $P_1 = P_2 = p$

Where, p can be found from the data,

$$p \quad \frac{n_1P_1 \quad n_2P_2}{n_1 \quad n_2}$$

$$q \quad 1 \quad p$$

p is the mean of the two proportions.

$$SE(P_1 \quad P_2) \quad \sqrt{pq \quad \frac{1}{n_1} \quad \frac{1}{n_2}}$$

$z \quad \dfrac{P_1 \quad P_2}{SE(P_1 \quad P_2)}$, P is approximately normal (0,1)

We write z ~ $N(0, 1)$

The usual rules for rejection or acceptance are applicable here.

• **Case (*II*):** If it is assumed that the proportion under question is not the same in the two populations from which the samples are drawn and that P_1, P_2 are the true proportions, we write,

$$SE(P_1 - P_2) = \sqrt{\left(\frac{P_1q_1}{n_1} + \frac{P_2q_2}{n_2}\right)}$$

We can also write the confidence interval for $P_1 - P_2$.

For two independent samples of sizes n_1, n_2 selected from two binomial populations, the 100 (1 – a) per cent confidence limits for $P_1 - P_2$ are,

$$(P_1 \quad P_2) \quad z_{/2}\sqrt{\frac{P_1q_1}{n_1} \quad \frac{P_2q_2}{n_2}}$$

The 90% confidence limits would be [with a = 0.1, 100 (1 – a) = 0.90]

$$(P_1 \quad P_2) \;\; 1.645\sqrt{\frac{P_1q_1}{n_1} \quad \frac{P_2q_2}{n_2}}$$

Consider Example 8.3 to further understand the test for equality.

Example 8.3: Out of 5000 interviewees, 2400 are in favour of a proposal, and out of another set of 2000 interviewees, 1200 are in favour. Is the difference significant?

Where, $\quad P_1 = \frac{2400}{5000} = 0.48 \qquad P_2 = \frac{1200}{2000} = 0.6$

Solution:

Given, $\quad P_1 = \frac{2400}{5000} = 0.48 \qquad P_2 = \frac{1200}{2000} = 0.6$

$n_1 = 5000 \qquad\qquad n_2 = 2000$

$$SE = \sqrt{\left(\frac{0.48 \times 0.52}{5000} + \frac{0.6 \times 0.4}{2000}\right)} = 0.013 \text{ (using Case (II))}$$

The difference is highly significant at 0.27 per cent level.

Large Sample Test for Equality of Two Means $\bar{X}_1, \bar{X}_2$

Suppose two samples of sizes n_1 and n_2 are drawn from populations having means μ_1, μ_2 and standard deviations σ_1, σ_2.

To test the equality of means $\bar{X}_1, \bar{X}_2$ we write,

$$H_0 : \;{}_1 \quad {}_2$$
$$H_1 : \;{}_1 \quad {}_2$$

If we assume H_0 is true, then

$$z \quad \frac{\bar{X}_1 \quad \bar{X}_2}{\sqrt{\frac{{}^2_1}{n_1} \quad \frac{{}^2_2}{n_2}}}$$, approximately normally distributed with mean 0, and S.D. = 1.

We write $z \sim N(0, 1)$

As usual, if $| z | > 2$ we reject H_0 at 4.55% level of significance, and so on (refer Example 8.4).

Example 8.4: Two groups of sizes 121 and 81 are subjected to tests. Their means are found to be 84 and 81 and standard deviations 10 and 12. Test for the significance of difference between the groups.

Solution:

$\bar{X}_1 = 84 \quad \bar{X}_2 = 81 \quad n_1 = 121 \quad n_2 = 81$

$\sigma_1 = 10 \quad \sigma_2 = 12$

$$z \quad \frac{\bar{X}_1 \quad \bar{X}_2}{\sqrt{\frac{{}^2_1}{n_1} \quad \frac{{}^2_2}{n_2}}}, \; z \quad \frac{84 \quad 81}{\sqrt{\overline{121} \quad \overline{81}}} = 1.86 < 1.96$$

The difference is not significant at the 5 per cent level of significance.

Small Sample Tests of Significance

$H \; : \; X$

$|t| \; {}^1 \frac{}{SE(\bar{X})}$

The sampling distribution of many statistics for large samples is approximately normal. For small samples with $n < 30$, the normal distribution, as shown in Example 8.4, can be used only if the sample is from a normal population with known σ.

If σ is not known, we can use student's t distribution instead of the normal. We then replace σ by sample standard deviation σ with some modification as shown.

Let $x_1, x_2, ..., x_n$ be a random sample of size n drawn from a normal population with mean μ and S.D. σ. Then,

$$t \quad \frac{\bar{x}}{s / \sqrt{n \quad 1}} \, .$$

Here, t follows the student's t distribution with $n - 1$ degrees of freedom.

Note: For small samples of $n < 30$, the term $\sqrt{n \quad 1}$, in SE = $s / \sqrt{n \quad 1}$, corrects the bias, resulting from the use of sample standard deviation as an estimator of σ.

Also,

$$\frac{s^2}{S^2} \quad \frac{n \;\; 1}{n} \text{ or } s \quad S\sqrt{\frac{n \quad 1}{n}}$$

Procedure: Small Samples

To test the null hypothesis H_0 : , against the alternative hypothesis

Calculate and compare it with the table value with $n - 1$ degrees of freedom

(d.f.) at level of significance 1 per cent.

NOTES

If this value > table value, reject H_0

If this value < table value, accept H_0

(Significance level idea same as for large samples)

We can also find the 95% (or any other) confidence limits for μ.

For the two-tailed test (use the same rules as for large samples; substitute t for z) the 95% confidence limits are,

$\bar{X} \quad t\, s / \sqrt{n \quad 1}$

Rejection Region

At a per cent level for two-tailed test if $|t| > t_{a/2}$ reject.

For one-tailed test, (right) if $t > t_a$ reject

(left) if $t > -t_a$ reject

At 5 per cent level the three cases are,

If $|t| > t_{0.025}$ reject two-tailed

If $t > t_{0.05}$ reject one-tailed right

If t £ $t_{0.05}$ reject one-tailed left

For proportions, the same procedure is to be followed.

Example 8.5: A firm produces tubes of diameter 2 cm. A sample of 10 tubes is found to have a diameter of 2.01 cm and variance 0.004. Is the difference significant? Given $t_{0.05,9}$= 2.26.

Solution:

$$t = \frac{\bar{X} - \mu}{s/\sqrt{n-1}}$$

$$= \frac{2.01-2}{\sqrt{0.004/(10-1)}}$$

$$= \frac{0.01}{0.021}$$

$$= 0.48$$

Since, $|t| < 2.26$, the difference is not significant at 5 per cent level.

3.4 CHI-SQUARE STATISTIC

Chi-square test is a non-parametric test of statistical significance for bivariate tabular analysis (also known as cross-breaks). Any appropriate test of statistical significance lets you know the degree of confidence you can have in accepting or rejecting a hypothesis. Typically, the Chi-square test is any statistical hypothesis test in which the test statistics has a chi-square distribution when the null hypothesis is true. It is performed on different samples (of people) who are different enough in some characteristic or aspect of their behaviour that we can generalize from the samples selected. The population from which our samples are drawn should also be different in the behaviour or characteristic. Amongst the several tests used in statistics for judging the significance of the sampling data, Chi-square test, developed by Prof. Fisher, is considered as an important test. Chi-square, symbolically written as χ^2 (pronounced as Ki-square), is a statistical

Check Your Progress

1. Define the term sampling distribution.
2. What does the Central Limit Theorem state?
3. What is a hypothesis?

measure with the help of which, it is possible to assess the significance of the difference between the observed frequencies and the expected frequencies obtained from some hypothetical universe. Chi-square tests enable us to test whether more than two population proportions can be considered equal. In order that Chi-square test may be applicable, both the frequencies must be grouped in the same way and the theoretical distribution must be adjusted to give the same total frequency which is equal to that of observed frequencies. χ^2 is calculated with the help of the following formula:

$$\chi^2 = \sum \frac{(f_0 - f_e)^2}{f_e}$$

Where, f_0 means the observed frequency; and

f_e means the expected frequency.

Whether or not a calculated value of χ^2 is significant, it can be ascertained by looking at the tabulated values of χ^2 for given degrees of freedom at a certain level of confidence (generally a 5 per cent level is taken). If the calculated value of χ^2 exceeds the table value, the difference between the observed and expected frequencies is taken as significant, but if the table value is more than the calculated value of χ^2, then the difference between the observed and expected frequencies is considered as insignificant, i.e., considered to have arisen as a result of chance and as such can be ignored.

Degrees of Freedom

The number of independent constraints determines the number of degrees of freedom (or *df*). If there are 10 frequency classes and there is one independent constraint, then there are (10 – 1) = 9 degrees of freedom. Thus, if n is the number of groups and one constraint is placed by making the totals of observed and expected frequencies equal, $df = (n - 1)$; when two constraints are placed by making the totals as well as the arithmetic means equal then $df = (n - 2)$, and so on. In the case of a contingency table (i.e., a table with two columns and more than two rows or table with two rows but more than two columns or a table with more than two rows and more than two columns) or in the case of a 2 × 2 table, the degrees of freedom is worked out as follows:

$$df = (c - 1)(r - 1)$$

Where, c = Number of columns

r = Number of rows

Conditions for the Application of Test

The following conditions should be satisfied before the test can be applied:

(i) Observations recorded and used are collected on a random basis.

(ii) All the members (or items) in the sample must be independent.

(iii) No group should contain very few items, say less than 10. In cases where the frequencies are less than 10, regrouping is done by combining the frequencies of adjoining groups so that the new frequencies become greater than 10. Some statisticians take this number as 5, but 10 is regarded as better by most of the statisticians.

(iv) The overall number of items (i.e., N) must be reasonably large. It should at least be 50, howsoever small the number of groups may be.

(v) The constraints must be linear. Constraints which involve linear equations in the cell frequencies of a contingency table (i.e., equations containing no squares or higher powers of the frequencies) are known as linear constraints.

NOTES

Areas of Application of Chi-Square Test

Chi-square test is applicable in large number of problems. The test is, in fact, a technique through the use of which it is possible for us to (a) Test the goodness of fit; (b) Test the homogeneity of a number of frequency distributions; and (c) Test the significance of association between two attributes. In other words, Chi-square test is a test of independence, goodness of fit and homogeneity. At times, Chi-square test is used as a test of population variance also.

As a test of goodness of fit, χ^2 test enables us to see how well the distribution of observed data fits the assumed theoretical distribution, such as Binomial distribution, Poisson distribution or the Normal distribution.

As a test of independence, χ^2 test helps explain whether or not two attributes are associated. For instance, we may be interested in knowing whether a new medicine is effective in controlling fever or not and χ^2 test will help us in deciding this issue. In such a situation, we proceed on the null hypothesis that the two attributes (viz., new medicine and control of fever) are independent. Which means that the new medicine is not effective in controlling fever. It may, however, be stated here that χ^2 is not a measure of the degree of relationship or the form of relationship between two attributes but it simply is a technique of judging the significance of such association or relationship between two attributes.

As a test of homogeneity, χ^2 test helps us in stating whether different samples come from the same universe. Through this test, we can also explain whether the results worked out on the basis of sample/samples are in conformity with well-defined hypothesis or the results fail to support the given hypothesis. As such, the test can be taken as an important decision-making technique.

As a test of population variance. Chi-square is also used to test the significance of population variance through confidence intervals, especially in case of small samples.

3.4.1 Additive Property of Chi-Square (χ^2)

An important property of χ^2 is its additive nature. This means that several values of χ^2 can be added together and if the degrees of freedom are also added, this number gives the degrees of freedom of the total value of χ^2. Thus, if a number of χ^2 values have been obtained from a number of samples of similar data, then, because of the additive nature of χ^2, we can combine the various values of χ^2 by just simply adding them. Such addition of various values of χ^2 gives one value of χ^2 which helps in forming a better idea about the significance of the problem under consideration. The following example illustrates the additive property of the χ^2 (refer Example 8.6).

Example 8.6: The following values of χ^2 are obtained from different investigations carried out to examine the effectiveness of a recently invented medicine for checking malaria.

Investigation	χ^2	*df*
1	2.5	1
2	3.2	1
3	4.1	1
4	3.7	1
5	4.5	1

What conclusion would you draw about the effectiveness of the new medicine on the basis of the five investigations taken together?

Solution:

By adding all the values of χ^2, we obtain a value equal to 18.0. Also, by adding the various d.f. as given in the question, we obtain a figure 5. We can now state that the value of χ^2 for 5 degrees of freedom (when all the five investigations are taken together) is 18.0.

Let us take the hypothesis that the new medicine is not effective. The table value of χ^2 for 5 degrees of freedom at 5% level of significance is 11.070. But our calculated value is higher than this table value which means that the difference is significant and is not due to chance. As such the hypothesis is wrong and it can be concluded that the new medicine is effective in checking malaria.

Important Characteristics of Chi-Square (χ^2) Test

The following are the important characteristics of chi-square test:

3.4.1.1 This test is based on frequencies and not on the parameters like mean and standarddeviation.

3.4.1.2 This test is used for testing the hypothesis and is not useful for estimation.

3.4.1.3 This test possesses the additive property.

3.4.1.4 This test can also be applied to a complex contingency table with several classes and as such is a very useful test in research work.

3.4.1.5 This test is an important non-parametric (or a distribution free) test as no rigid assumptions are necessary in regard to the type of population and no need of the parameter values. It involves less mathematical details.

A Word of Caution in Using χ^2 Test

Chi-square test is no doubt a most frequently used test, but its correct application is equally an uphill task. It should be borne in mind that the test is to be applied only when the individual observations of sample are independent which means that the occurrence of one individual observation (event) has no effect upon the occurrence of any other observation (event) in the sample under consideration. The researcher, while applying this test, must remain careful about all these things and must thoroughly understand the rationale of this important test before using it and drawing inferences concerning his hypothesis.

3.5 *t*-STATISTIC

Sir William S. Gosset (pen name Student) developed a significance test and through it made significant contribution to the theory of sampling applicable in case of small samples.

When population variance is not known, the test is commonly known as Student's *t*-test and is based on the *t* distribution.

Like the normal distribution, *t* distribution is also symmetrical but happens to be flatter than the normal distribution. Moreover, there is a different *t* distribution for every possible sample size. As the sample size gets larger, the shape of the *t* distribution loses its flatness and becomes approximately equal to the normal distribution. In fact, for sample sizes of more than 30, the *t* distribution is so close to the normal distribution that we will use the normal to approximate the *t* distribution. Thus, when *n* is small, the *t* distribution is far from normal, but when *n* is infinite, it is identical to normal distribution.

For applying *t*-test in context of small samples, the *t* value is calculated first of all and, then the calculated value is compared with the table value of *t* at certain level of significance for given degrees of freedom. If the calculated value of *t* exceeds the table value (say $t_{0.05}$), we infer that the difference is significant at 5 per cent level, but if the calculated value is t_0, is less than its concerning table value, the difference is not treated as significant.

The *t*-test is used when the following two conditions are fullfiled:

(i) The sample size is less than 30, i.e., when $n \leq 30$.

(ii) The population standard deviation (σ_p) must be unknown.

In using the *t*-test, we assume the following:

(i) The population is normal or approximately normal.

(ii) The observations are independent and the samples are randomly drawn samples.

(iii) There is no measurement error.

(iv) In the case of two samples, population variances are regarded as equal if equality of the two population means is to be tested.

The following formulae are commonly used to calculate the *t* value:

(i) To Test the Significance of the Mean of a Random Sample

Where, $\bar{X}$ = Mean of the sample

μ = Mean of the universe

= S.E. of mean in case of small sample and is worked out as,

and the degrees of freedom = $(n - 1)$

The above stated formula for *t* can as well be stated as,

$$t = \frac{|\bar{x} - \mu|}{SE_{\bar{x}}}$$

$$= \frac{|\bar{x} - \mu|}{\dfrac{\dfrac{\sqrt{\Sigma(x-\bar{x})^2}}{n-1}}{\sqrt{n}}}$$

$$= \frac{|\bar{x} - \mu|}{\sqrt{\dfrac{\Sigma(x-\bar{x})^2}{n-1}}} \times \sqrt{n}$$

If we want to work out the probable or fiducial limits of population mean (μ) in case of small samples, we can use either of the following:

(a) Probable limits with 95 per cent confidence level:

$$\mu = \bar{X} \pm SE_{\bar{x}}(t_{0.05})$$

(b) Probable limits with 99 per cent confidence level:

$$\mu = \bar{X} \pm SE_{\bar{x}}(t_{0.01})$$

At other confidence levels, the limits can be worked out in a similar manner, taking the concerning table value of t just as we have taken $t_{0.05}$ in (a) and $t_{0.01}$ in (b) above.

(ii) To Test the Difference between the Means of Two Samples

$$t = \frac{|\bar{X}_1 - \bar{X}_2|}{\text{SE}_{\bar{x}_1 - \bar{x}_2}}$$

Where, $\bar{X}_1$ = Mean of the sample 1

$\bar{X}_2$ = Mean of the sample 2

$SE_{\bar{x}_1-\bar{x}_2}$ = Standard error of difference between two sample means and is worked out as follows:

$$SE_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{\sum(X_{1i} - \bar{x}_1)^2 + \sum(X_{2i} - \bar{x}_2)^2}{n_1 + n_2 - 2}} \times \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

and the degrees of freedom = $(n_1 + n_2 - 2)$.

When the actual means are in fraction, then use of assumed means is convenient. In such a case, the standard deviation of difference, i.e.,

$$\sqrt{\frac{\Sigma(x_{1i} + x_1)^2 + \Sigma(x_{2i} - \bar{x}_2)^2}{n_1 + n_2 - 2}}$$

can be worked out by the following short-cut formula:

$$= \frac{\sqrt{\Sigma(x_{1i} - A_1)^2} + \Sigma(x_{2i} - A_1)^2 - n_1(x_{1i} - A_2)^2 - n_2(x_{2i} - A_2)^2}{n_1 + n_2 - 2}$$

NOTES

Where, A_1 = Assumed mean of sample 1

A_2 = Assumed mean of sample 2

X_1 = True mean of sample 1

X_2 = True mean of sample 2

(iii) To Test the Significance of an Observed Correlation Coefficient

$$t = \frac{r}{\sqrt{1-r^2}} \times \sqrt{n-2}$$

Here, t is based on $(n-2)$ degrees of freedom.

(iv) In Context of the 'Difference Test'

Difference test is applied in the case of paired data and in this context t is calculated as,

$$t = \frac{\bar{x}_{Ditt} - 0}{6_{Diff}\sqrt{n}} = \frac{\bar{x}_{Diff} - 0}{6_{Diff}}\sqrt{n}$$

Where, $\bar{X}_{Diff}$ or $\bar{D}$ = Mean of the differences of sample items.

0 = the value zero on the hypothesis that there is no difference

$\sigma_{Diff.}$ = standard deviation of difference and is worked out as

$$\sqrt{\frac{\sum D - \bar{X}_{Diff})^2}{(n-1)}}$$

or

$$\sqrt{\frac{\Sigma D^2 - (\bar{D})^2 n}{(n-1)}}$$

D = differences

n = number of pairs in two samples and is based on $(n-1)$ degrees of freedom

3.6 *F*-STATISTIC

In business decisions, we are often involved in determining if there are significant differences among various sample means, from which conclusions can be drawn about the differences among various population means. What if we have to compare more than two sample means? For example, we may be interested to find out if there are any significant differences in the average sales figures of four different salesmen employed by the same company, or we may be interested to find out if the average monthly expenditures of a family of 4 in 5 different localities are similar or not, or the telephone company may be interested in checking, whether there are any significant differences in the average number of requests for information received in a given day among the five areas of New York City, and so on. The methodology used for such types of determinations is known as Analysis of Variance.

NOTES

This technique is one of the most powerful techniques in statistical analysis and was developed by R.A. Fisher. It is also called the *F*-Test.

There are two types of classifications involved in the analysis of variance. The one-way analysis of variance refers to the situations when only one fact or variable is considered. For example, in testing for differences in sales for three salesman, we are considering only one factor, which is the salesman's selling ability. In the second type of classification, the response variable of interest may be affected by more than one factor. For example, the sales may be affected not only by the salesman's selling ability, but also by the price charged or the extent of advertising in a given area.

For the sake of simplicity and necessity, our discussion will be limited to One-way Analysis of Variance (ANOVA).

The null hypothesis, that we are going to test, is based upon the assumption that there is no significant difference among the means of different populations. For example, if we are testing for differences in the means of *k* populations, then,

$$H_0 = \mu_1 = \mu_2 = \mu_3 = \ldots\ldots.= \mu_k$$

The alternate hypothesis (H_1) will state that at least two means are different from each other. In order to accept the null hypothesis, all means must be equal. Even if one mean is not equal to the others, then we cannot accept the null hypothesis. The simultaneous comparison of several population means is called *Analysis of Variance or ANOVA.*

Assumptions

The methodology of ANOVA is based on the following assumptions:

(i) Each sample of size n is drawn randomly and each sample is independent of the other samples.

(ii) The populations are normally distributed.

(iii) The populations from which the samples are drawn have equal variances. This means that:

$$\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \ldots\ldots = \sigma_k^2,$$

for k populations.

The Rationale Behind Analysis of Variance

Why do we call it the Analysis of Variance, even though we are testing for means? Why not simply call it the Analysis of Means? How do we test for means by analysing the variances? As a matter of fact, in order to determine if the means of several populations are equal, we do consider the measure of variance, σ^2.

The estimate of population variance, σ^2, is computed by two different estimates of σ^2, each one by a different method. One approach is to compute an estimator of σ^2 in such a manner that even if the population means are not equal, it will have no effect on the value of this estimator. This means that, the differences in the values of the population means do not alter the value of σ^2 as calculated by a given method. This estimator of σ^2 is the average of the variances found within each of the samples. For example, if we take 10 samples of size *n*, then each sample will have a mean and a variance. Then, the mean of these 10 variances would be considered as an unbiased estimator of σ^2, the population variance, and its value remains appropriate irrespective of whether the population means are equal or not. This is really done by pooling all the sample variances

NOTES

to estimate a common population variance, which is the average of all sample variances. This common variance is known as variance within samples or σ^2_{within}.

The second approach to calculate the estimate of σ^2, is based upon the Central Limit Theorem and is valid only under the null hypothesis assumption that all the population means are equal. This means that in fact, if there are *no differences* among the population means, then the computed value of σ^2 by the second approach should not differ significantly from the computed value of σ^2 by the first approach.

Hence,

If these two values of σ^2 are approximately the same, then we can decide to accept the null hypothesis.

The second approach results in the following computation:

Based upon the Central Limit Theorem, we have previously found that the standard error of the sample means is calculated by,

$$\sigma^2_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

or, the variance would be:

$$\sigma^2_{\bar{x}} = \frac{\sigma^2}{n}$$

or, $$\sigma^2 = n\sigma^2_{\bar{x}}$$

Thus, by knowing the square of the standard error of the mean , we could multiply it by n and obtain a precise estimate of σ^2. This approach of estimating σ^2 is known as $\sigma^2_{\text{between}}$. Now, if the null hypothesis is true, that is if all population means are equal then, $\sigma^2_{\text{between}}$ value should be approximately the same as σ^2_{within} value. A significant difference between these two values would lead us to conclude that this difference is the result of differences between the population means.

But, how do we know that any difference between these two values is significant or not? How do we know whether this difference, if any, is simply due to random sampling error or due to actual differences among the population means?

R.A. Fisher developed a Fisher test or F-test to answer the above question. He determined that the difference between $\sigma^2_{\text{between}}$ and σ^2_{within} values could be expressed as a ratio to be designated as the F-value, so that,

In the minters case, if the population means are exactly the same, then $\sigma^2_{\text{between}}$ will be equal to the σ^2_{within} and the value of F will be equal to 1.

However, because of sampling errors and other variations, some disparity between these two values will be there, even when the null hypothesis is true, meaning that all population means are equal. The extent of disparity between the two variances and consequently, the value of F, will influence our decision on whether to accept or reject the null hypothesis. It is logical to conclude that, if the population means are not equal, then their sample means will also vary greatly from one another, resulting in a larger value of $\sigma^2_{\text{between}}$ and hence a larger value of F (σ^2_{within} is based only on sample variances and not on sample means and hence, is not affected by differences in sample means). Accordingly, the larger the value of F, the more likely the decision to reject the null hypothesis. But, how large the value of F be so as to reject the null hypothesis? The answer is that the computed value of F must be larger than the *critical* value of F, given in the table for a given level of significance and calculated number of degrees of freedom. (The F distribution is a family of curves, so that there are different curves for different degrees of freedom).

Degrees of Freedom

We have talked about the F-distribution being a family of curves, each curve reflecting the degrees of freedom relative to both $\sigma^2_{between}$ and σ^2_{within}. This means that, the degrees of freedom are associated both with the numerator as well as with the denominator of the F-ratio.

(i) **The numerator.** Since the variance between samples, $\sigma^2_{between}$ comes from many samples and if there are k number of samples, then the degrees of freedom, associated with the numerator would be $(k-1)$.

(ii) **The denominator** is the *mean variance* of the variances of k samples and since, each variance in each sample is associated with the size of the sample (n), then the degrees of freedom associated with each sample would be $(n - 1)$. Hence, the total degrees of freedom would be the sum of the degrees of freedom of k samples or

$df = k(n-1)$, when each sample is of size n.

The F-Distribution

The major characteristics of the F-distribution are as follows:

(i) Unlike normal distribution, which is only one type of curve irrespective of the value of the mean and the standard deviation, the F distribution is a *family* of curves. A particular curve is determined by two parameters. These are the degrees of freedom in the numerator and the degrees of freedom in the denominator. The shape of the curve changes as the number of degrees of freedom changes.

(ii) It is a continuous distribution and the value of F cannot be negative.

(iii) The curve representing the F distribution is positively skewed.

(iv) The values of F theoretically range from zero to infinity.

A diagram of F distribution curve is shown in Figure 8.1.

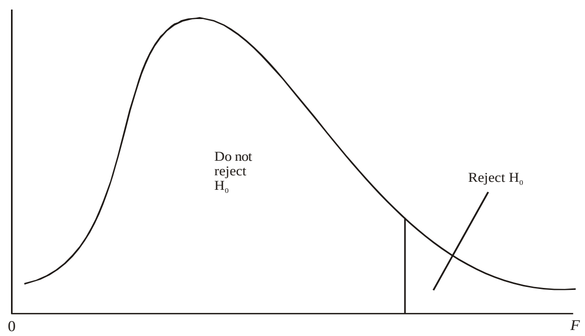


Fig. 8.1 *F-Distribution on Curve*

The rejection region is only in the right end tail of the curve because unlike Z distribution and t distribution which had negative values for areas below the mean, F distribution has only positive values by definition and only positive values of F that are larger than the critical values of F, will lead to a decision to reject the null hypothesis.

NOTES

Computation of *F*

F ratio contains only two elements, which are the variance between the samples and the variance within the samples.

If all the means of samples were exactly equal and all samples were exactly representative of their respective populations so that all the sample means were exactly equal to each other and to the population mean, then there will be no variance. However, this can never be the case. We always have variation, both between samples and within samples, even if we take these samples randomly and from the same population. This variation is known as the total variation.

The total variation designated by where X represents individual observations

for all samples and is the grand mean of all sample means and equals (μ), the population mean, is also known as the *total sum of squares* or *SST*, and is simply the sum of squared differences between each observation and the overall mean. This total variation represents the contribution of two elements. These elements are:

(i) ***Variance between Samples:*** The variance between samples may be due to the effect of different *treatments*, meaning that the population means may be affected by the *factor* under consideration, thus making the population means actually different, and some variance may be due to the inter-sample variability. This variance is also known as the sum of squares between samples. Let this sum of squares be designated as *SSB*.

Then, *SSB* is calculated by the following steps:

(a) Take k samples of size n each and calculate the mean of each sample, i.e.,

(b) Calculate the grand mean of the distribution of these sample means, so that,

(c) Take the difference between the means of the various samples and the grand mean, i.e.,

(d) Square these deviations or differences individually, multiply each of these squared deviations by its respective sample size and sum up all these products, so that we get;

$$\sum_{i=1}^{k} n_i(\bar{X}_i - \bar{\bar{X}})^2, \text{ where } n_i = \text{size of the } i\text{th sample.}$$

This will be the value of the *SSB*.

However, if the individual observations of all samples are not available, and only the various means of these samples are available, where the samples are either of the same size n or different sizes, n_i, n_2, n_3, , n_k, then the value of SSB can be calculated as:

$$SSB = n_i(\bar{X}_i - \bar{\bar{X}})^2 + n_2(\bar{X}_2 - \bar{\bar{X}})^2 + \; ..... \; n_k(\bar{X}_k - \bar{\bar{X}})^2$$

Where,

n_1 = Number of items in sample 1

n_2 = Number of items in sample 2

n_k = Number of items in sample k

$\bar{X}_1$ = Mean of sample 1

$\bar{X}_2$ = Mean of sample 2

$\bar{X}_k$ = Mean of sample k

$\bar{\bar{X}}$ = Grand mean or average of all items in all samples.

(e) Divide SSB by the degrees of freedom, which are $(k - 1)$, where k is the number of samples and this would give us the value of $\sigma^2_{\text{between}}$, so that,

$$\sigma^2_{\text{between}} = \frac{SSB}{(k-1)}$$

(This is also known as mean square between samples or *MSB)*.

(ii) *Variance within Samples:* Even though each observation in a given sample comes from the same population and is subjected to the same treatment, some chance variation can still occur. This variance may be due to sampling errors or other natural causes. This variance or sum of squares is calculated by the following steps:

(a) Calculate the mean value of each sample, i.e., $\bar{X}_1, \bar{X}_2, \bar{X}_3, ... \bar{X}_k$.

(b) Take one sample at a time and take the deviation of each item in the sample from its mean. Do this for all the samples, so that we would have a difference between each value in each sample and their respective means for all values in all samples.

(c) Square these differences and take the total of all these squared differences (or deviations). This sum is also known as *SSW* or sum of squares within samples.

(d) Divide this *SSW* by the corresponding degrees of freedom. The degrees of freedom are obtained by subtracting the total number of samples from the total number of items. Thus, if N is the total number of items or observations, and k is the number of samples, then,

$$df = (N - k)$$

These are the degrees of freedom within samples. (If all samples are of equal size n, then $df = k(n - 1)$, since $(n - 1)$ are the degrees of freedom for each sample and there are k samples).

(e) This figure SSW/df, is also known as σ^2_{within}, or MSW (mean of sum of squares within samples).

Now, the value of F can be computed as:

NOTES

This value of F is then compared with the critical value of F from the table and a decision is made about the validity of null hypothesis.

3.7 ONE-TAILED AND TWO-TAILED TESTS

A one-tailed test requires rejection of the null hypothesis when the sample statistic is greater than the population value or less than the population value at a certain level of significance.

1. We may want to test if the sample mean exceeds the population mean m. Then the null hypothesis is,

   H_0: > μ

2. In the other case the null hypothesis could be,

   H_0: < μ

Each of these two situations leads to a one-tailed test and has to be dealt with in the same manner as the two-tailed test. Here the critical rejection is on one side only, right for > μ and left for < μ. Both the Figures 8.2 and 8.3 here show a five per cent level of test of significance.

For example, a minister in a certain government has an average life of 11 months without being involved in a scam. A new party claims to provide ministers with an average life of more than 11 months without scam. We would like to test if, on the average, the new ministers last longer than 11 months. We may write the null hypothesis H_0: = 11 and alternative hypothesis H_1: > 11.

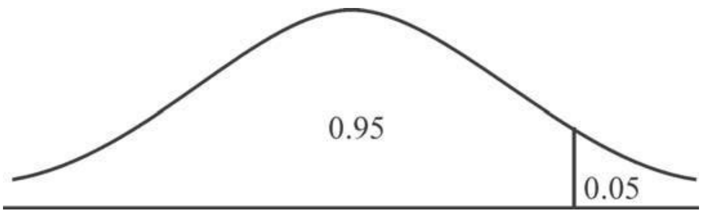


Fig. 8.2 H_0: $\bar{X}$ > μ

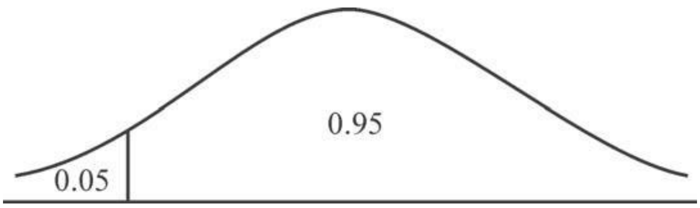


Fig. 8.3 H_0: < μ

8.7.1 One Sample Test

So far we have discussed situations in which the null hypothesis is rejected if the sample statistic $\bar{X}$ is either too far above or too far below the population parameter μ, which means that the area of rejection is at both ends (or tails) of the normal curve. For example, if we are testing for the average IQ of the college students being equal to 130, then the null hypothesis $H_0 : \mu = 130$ will be rejected if a sample selected gives a mean which is either too high or too low compared to μ. This can be expressed as follows:

$$H_0 : \mu = 130$$
$$H_1 : \mu \neq 130$$

This means that with α = 0.05 (95% confidence interval), the value of must be within of the assumed value of μ under H_0 in order to accept the null hypothesis. In other words, must be less than ± 1.96. The element

$\frac{\bar{X} - \mu\,(\text{under } H_0)}{\sigma_{\bar{X}}}$ is known as the critical ratio or CR. It means that:

At α = 0.05, accept H_0 if critical ratio CR falls within ± 1.96 and reject H_0 if CR is less than (–1.96) or greater than (+1.96). If it happens to be exactly 1.96 then we can accept H_0.

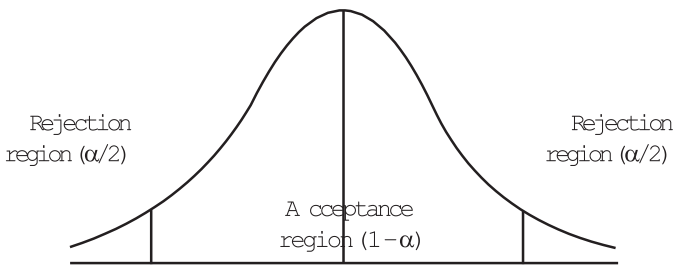
On the other hand, there are situations in which the area of rejection lies entirely on one extreme of the curve, which is either the right end of the tail or the left end of the tail. Tests concerning such situations are known as *one-tailed* tests, and the null hypothesis is rejected only if the value of the sample statistic falls into this single rejection region.

$\bar{X}$ 1.96 $\sigma_{\bar{X}}$ $\frac{\bar{X} - \mu\,(\text{under } H_0)}{\sigma_{\bar{X}}}$

For example, let us assume that we are manufacturing 9 volt batteries and we claim that our batteries last on an average (μ) 100 hours. If somebody wants to test the accuracy of our claim, he can take a random sample of our batteries and find the average ($\bar{X}$) of this sample. He will reject our claim only if the value of so calculated is considerably lower than 100 hours, but will not reject our claim if the value of is considerably higher than 100 hours. Hence in this case, the rejection area will only be on the left end tail of the curve.

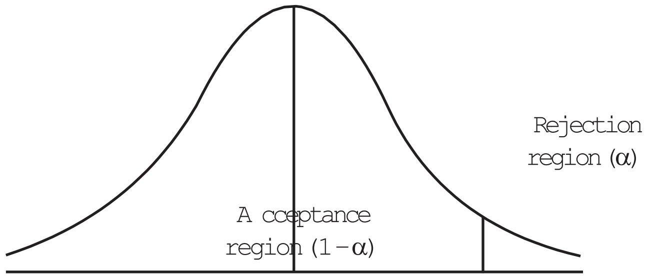
Similarly, if we are making a low calorie diet ice cream and claim that it has on an average only 500 calories per pound and an investigator wants to test our claim, he can take a sample and compute . If the value of is much higher than 500 calories, then he will reject our claim. But he will not reject our claim if the value of is much lower than 500 calories. Hence the rejection region in this case will be only on the right end tail of the curve. These rejection and acceptance areas are shown in the normal curves as follows:

(a) Two-Tailed Test

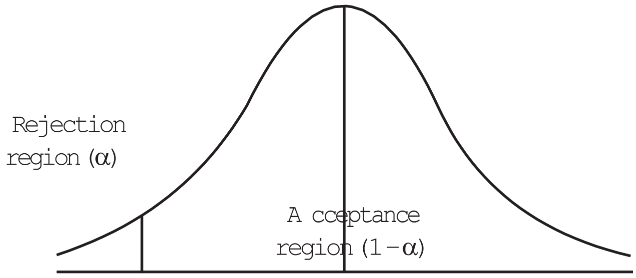


NOTES

(b) Right-Tailed Test



(c) Left-Tailed Test



Tests Involving a Population Mean (Large Sample)

This type of testing involves decisions to check whether a reported population mean is reasonable or not, compared to the sample mean computed from the sample taken from the same population. A random sample is taken from the population and its statistic is computed. An assumption is made about the population mean μ as being equal to the sample mean and a test is conducted to see if the difference (– μ) is significant or not. This difference is not significant if it falls within the acceptance region and this difference is considered significant if it falls within the rejection region or the critical region at a given level of significance α.

It must also be noted that if population is not known to be normally distributed, then the sample size should be large enough, generally more than 30. However, if population is known to be normally distributed and the population standard deviation is known then even a smaller sample size would be acceptable.

Example 8.7: (Two-Tailed Test)

Assume that the average annual income for government employees in the nation is reported by the Census Bureau to be $18,750.00. There was some doubt whether the average yearly income of government employees in Washington was representative of the national average.

A random sample of 100 government employees in Washington was taken and it was found that their average salary was $19,240.00 with a standard deviation of $2,610.00. At a level of significance α = 0.05 (95% confidence level), can we conclude that the average salary of government employees in Washington is representative of the national average?

Solution: Obviously, it is a two-tailed test because if the salary of government employees in Washington is too high or too low compared to the national average then the hypothesis that the average salary of government employees in Washington is no different than the national average would be rejected.

Following the steps described in the procedure for hypothesis testing, we find:

1. Null hypothesis: $H_0 : \mu = \$18,750$

   Alternate hypothesis: $H_1 : \mu \neq \$18,750$
2. Level of significance as given $\alpha = 0.05$.
3. Determination of a suitable test statistic. Since we are testing for the population mean and according to the Central Limit Theorem, the sampling distribution of the sample means is approximately normally distributed with a standard error of the mean being $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$, the following test statistic would be appropriate:

$$Z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}}$$

Hence,
$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

Where,

$\bar{x}$ = Sample mean

μ = Population mean

σ = Standard deviation of the population (s)

(Since population standard deviation is not given and the sample size is large enough, we can approximate the sample standard deviation s as equivalent to population standard deviation σ.)

4. Defining the critical region. Since $\alpha = 0.05$ and it is a two-tailed test, the rejection region will be on both end tails of the curve in such a way that the rejection area will comprise 2.5% at the end of the right tail and 2.5% at the end of the left tail. In other words, at $\alpha = 0.05$, the region of acceptance is enclosed by the value of Z being $\pm$ 1.96 around the mean.

Now for our example, let us calculate the value of Z.

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

Where,

$\bar{x}$ = 19240

μ = 18750

σ = s = 2610

n = 100

Then,
$$Z = \frac{19240 - 18750}{2610/\sqrt{100}}$$

$$= \frac{490}{261} = 1.877$$

Now, since the calculated value of Z as 1.877 is less than 1.96 and falls within the area of acceptance bounded by Z = ±1.96, we cannot reject the null hypothesis.

We could also solve this problem by constructing a 95% confidence interval for the population mean and then testing whether the sample mean falls within the confidence interval. The confidence interval is bounded by μ , as illustrated below:

Now,	$X_1 = \mu - Z\sigma_{\bar{x}}$
and	$X_2 = \mu + Z$
We know that,	$\mu = 18750$
	$Z = 1.96$ at $\alpha = 0.05$
	=
Then,	$X_1 = 18750 - 1.96(261)$
	$= 18238.44$
and	$X_2 = 18750 + 1.96(261)$
	$= 19261.56$

This means that if the sample mean lies within these two limits, then we cannot reject the null hypothesis. As we can see, the sample mean of $19,240.00 lies within this interval, so that we cannot reject the null hypothesis. Hence, our decision is to conclude that there is no significant difference between the average salary of government employees in Washington and the national average and it is purely coincidental that the average salary of government employees in Washington is numerically different than the national average.

Example 8.8: One-Tailed Test (Left-Tail)

The manufacturer of light bulbs claims that a light bulb lasts on an average 1600 hours. We want to test his claim. We will not reject his claim if the average of the sample taken lasts considerably more than 1600 hours, built we will reject his claim if it lasts considerably less than 1600 hours. Hence, it is a one-tailed test and the area of rejection is the left end tail of the curve.

A sample of 100 light bulbs was taken at random and the average bulb life of this sample was computed to be 1570 hours with a standard deviation of 120 hours. At $\alpha = 0.01$, let us test the validity of the claim of this manufacturer.

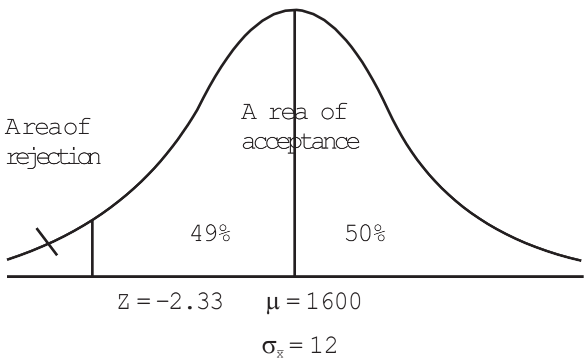
NOTES

Solution: Since the sample is large ($n = 100$), we can approximate the population standard deviation (σ) by sample standard deviation (s) so that:

Null hypothesis: $H_0 : \mu = 1600$

Alternate hypothesis : $H_1 : \mu < 1600$

Then at 99% confidence interval ($\alpha = 0.01$), the acceptance region is bounded by Z = –2.33 on the left tail of the standardized normal curve as shown below:



Now, $$Z = \frac{\bar{X} - \mu}{\sigma_{\bar{x}}}$$

Where, $\bar{X} = 1570$

$\mu = 1600$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{s}{\sqrt{n}} = \frac{120}{\sqrt{100}} = 12$$

Then: $Z = \frac{1570 - 1600}{12} = -\left(\frac{30}{12}\right) = -(2.5)$

Since our computed value of Z is numerically larger than the critical value of Z which is – (2.33), we cannot accept the null hypothesis at 99% confidence interval. (The negative sign is simply a concept that the value lies on the left of the mean μ and it is not an algebraic sign.) This means that the manufacturer's claim is not valid.

Example 8.9: One-Tailed Test (Right Tail)

An insurance company claims that it takes 2 weeks (14 days), on an average, to process an auto accident claim. The standard deviation is 6 days. To test the validity of this claim, an investigator randomly selected 36 people who recently filed claims. This sample revealed that it took the company an average of 16 days to process these claims. At 99% level of confidence, check if it takes the company more than 14 days on an average to process a claim.

Solution: In this case, the population parameter being tested is μ which is the average number of days it takes the company to process a claim. The company's claim is not valid if it takes considerably longer than the 14 days it claims on an average to process a claim. Hence,

$H_0 : \mu = 14$

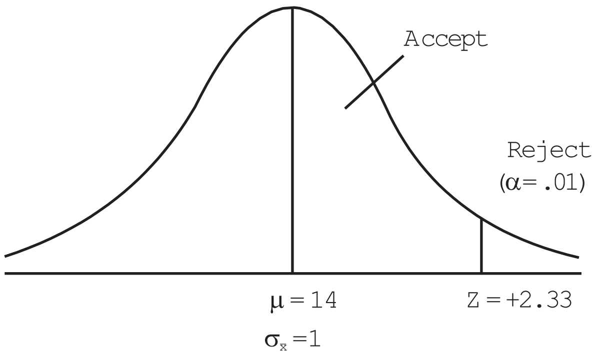
$H_1 : \mu > 14$

NOTES

Then,

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

$$Z =$$



Since the Z value of 2 is less than the critical value of Z which is 2.33, and it falls within the region of acceptance, we cannot reject the null hypothesis. Accordingly the company's claim is considered to be valid.

Tests Involving A Single Proportion

So far, we have dealt with the population parameter μ which reflects quantitative data. It cannot be used for qualitative data. For such qualitative data, the parameter of interest is the population proportion favouring one of the outcomes of the event. There are many situations in which we must test the validity of statements about the population proportions or percentages. For example, if a politician claims that 60% of the population supports his viewpoint on a given issue, we can test this claim by taking random samples of people and asking their opinions about this politician and finding the percentage of people on an average who support the viewpoint of this politician and then test whether this sample percentage is significantly different than his claim of population percentage. This technique is used in analysing the qualitative data where we can test for the presence or absence of a certain characteristic. For example, we may want to know if the government figures on the unemployment situation are accurate or not. Suppose that the government figures indicate that 9% of the work force is unemployed. We can always take a random sample and check for its validity.

This type of data follows the binomial distribution with:

Sample proportion $p =$

However, if n is large enough, so that,

$$np \geq 5$$

and $$n(1-p) \geq 5$$

Then it can be approximated to normal distribution and test statistic Z can be used.

Where,

Z =

NOTES

π = Population proportion

p = Sample proportion

σ_P=

Then the computed value of Z is compared with the critical value of Z in order to accept or reject the null hypothesis.

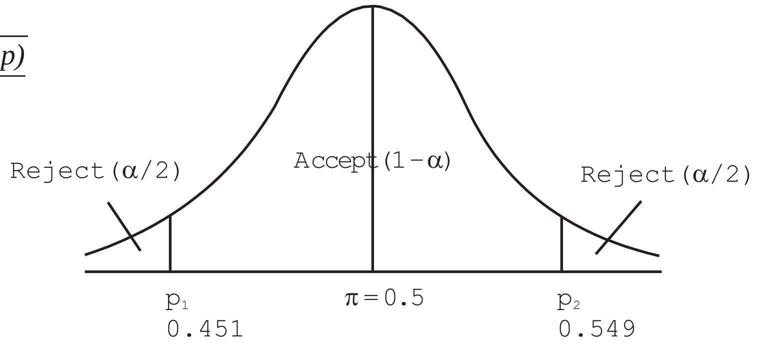
The testing of hypothesis follows the same procedure as in the case of tests about the population means and can best be illustrated with the help of following example.

Example 8.10: The sponsor of a television show believes that his studio audience is divided equaly between men and women. Out of 400 persons attending the show one day, there where 230 men. At $\alpha = 0.05$, test if the belief of the sponsor is correct.

Solution: This is a two-tailed test, since too many men as well as too few men in the audience would become the cause of rejection of the null hypothesis.

In order for the null hypothesis to be accepted, the sample proportion $p = (x/n)$ must fall within the confidence interval bounded by p_1 and p_2 as shown in the diagram which is the area of acceptance.

$\sqrt{\frac{p\ \pi\ (1\pi - \pi)}{\sigma_p\ n}}$ or $= \sqrt{\frac{p(1-p)}{n}}$



Here,

Null hypothesis: $H_0 : p = 0.5$

Alternate hypothesis: $H_1 : p \neq 0.5$

Confidence interval is defined as follows:

$$p_1 = p - Z\sigma_p$$
$$p_2 = p + Z\sigma_p$$

Where,

$$\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}} = \sqrt{\frac{(0.5)(0.5)}{400}} = 0.025$$

$Z = 1.96$ (at $\alpha = 0.05$)

$\pi = 0.5$

NOTES

Substituting these values we get,

$$p_1 = 0.5 - 1.96(0.025) = 0.451$$

$$p_2 = 0.5 + 1.96(0.025) = 0.549$$

In our example, the sample proportion $p = x/n = 230/400 = 0.575$. Clearly our sample proportion lies outside the region of acceptance and is in the critical region. Hence the null hypothesis cannot be accepted.

An alternate method to test the validity of the null hypothesis would be to compute the value of Z for the given information and compare it with the critical value of Z from the table, which, at 0.05 level of significance is 1.96.

Now,

$$Z = \frac{p - \pi}{\sigma_p}$$

Where,

p = Sample proportion = 230/400 = 0.575

π = Population proportion = 0.5

$\sigma_p =$ = 0.025 (as calculated above).

Then,

$$Z = \frac{0.575 - 0.500}{0.025}$$

$$= \frac{0.075}{0.025} = 3$$

Since our computed value of Z = 3, is higher than the critical value of Z = 1.96, we cannot accept the null hypothesis.

Example 8.11: One-Tailed Test

The mayor of the city claims that 60% of the people of the city follow him and support his policies. We want to test whether his claim is valid or not. A random sample of 400 persons was taken and it was found that 220 of these people supported the mayor. At level of significance $\alpha = 0.01$ what can we conclude about the mayor's claim.

Solution: Clearly, it is a one-tailed test for we will only reject the mayor's claim if the sample proportion of persons who support the mayor is considerably less than the mayor's claim about the population proportion of persons who support him. We will not reject his claim if such sample proportion is considerably higher than the population proportion.

Hence:

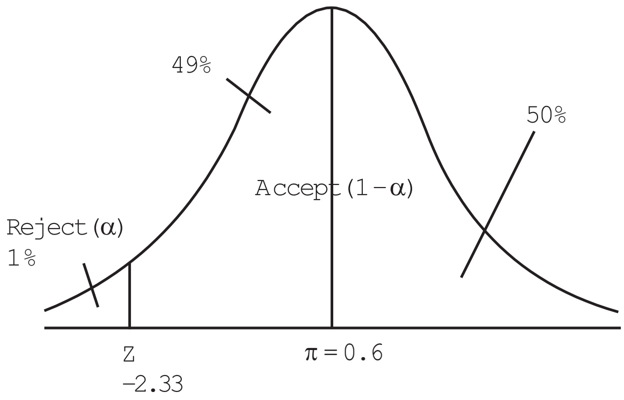
Null hypothesis: $H_0 : \pi = 0.6$

Alternate hypothesis: $H_1 : \pi < 0.6$

(The null hypothesis may also be expressed as $H_0 : \pi \geq 0.6$).

The sample proportion $p = 220/400 = 0.55$

NOTES



Now,

$$\sigma_p =$$

$$= \sqrt{\frac{(0.6)(0.4)}{400}}$$

$$= 0.0245$$

Then,

$$Z = \frac{p - \pi}{\sigma_p}$$

$$=$$

$$= -(2.04)$$

Since, it is a one-tailed test, the critical value of Z = – (2.33) for α = 0.1. Ignoring the negative sign, we note that the numerical value of our computed Z is less than the numerical critical value of Z, and hence we cannot reject the null hypothesis.

3.7.2 Two Sample Test for Large Samples

In many decision-making situations, comparison of two population means or two population proportions, becomes an area of interest. For example, we may be interested in comparing the effectiveness of two different teaching methods, where the effectiveness would be measured by the difference in the average student achievement under the two different techniques. Or, we may be interested to know if there is any significant difference in the average age of life for men and women in this country. Or, we may be interested to know if the average expenditure of two different communities are significantly different from each other. For this purpose, we can test one population mean against the other and draw conclusions for the purpose of making rational decisions.

NOTES

Testing the Difference between Two Sample Means

So far we have discussed sampling distribution of the means where a hypothesis was tested for any significant difference between the sample mean and the population mean. Now, we are interested to know if there are any significant differences between two population means. Let us assume that we want to find out if there is any significant difference in the average age of students who graduate with a bachelor degree in business from Baruch college and from Medgar Evers college. We take corresponding samples of graduating seniors from both colleges and find the mean of each sample taken from each college. Let these means and the differences in these means be represented as follows:

Baruch(1)	*Medgar Evers (2)*	*Differences*
X_{11}	X_{21}	$X_{11} - X_{21}$
X_{12}	X_{22}	$\bar{X}_{12} - X_{22}$
X_{13}	X_{23}	$\bar{X}_{13} - X_{23}$
X_{14}	X_{24}	$\bar{X}_{14} - X_{24}$
•	•	•
•	•	•
•	•	•
X_{1n}	X_{2n}	$\bar{X}_{1n} - X_{2n}$

In the above example, the first subscript represents the college and the second subscript represents the sequential sample number.

Now we have a distribution of the differences in the sample means. This is known as the sampling distribution of $(\bar{X}_1 - \bar{X}_2)$.

Basing on our analysis of the Central Limit Theore, we can make the following statements concerning the sampling distribution of the difference between sample means

.

If two independent samples of size n_1 and n_2 (both n_1 and n_2 to be larger than 30) are taken from populations with mean μ_1 and μ_2, and standard deviation σ_1 and σ_2, distribution with the following properties,

(a) The mean of the sampling distribution of is $(\mu_1 - \mu_2)$.

(b) The standard error of differences of sample means is given by,

However, if σ_1 and σ_2 are not known, then since n_1 and n_2 are sufficiently l;arge, the standard error of this distribution can be approximated by,

$$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

For the purpose of testing the hypothesis, we can use the standard normal distribution to find the Z score as,

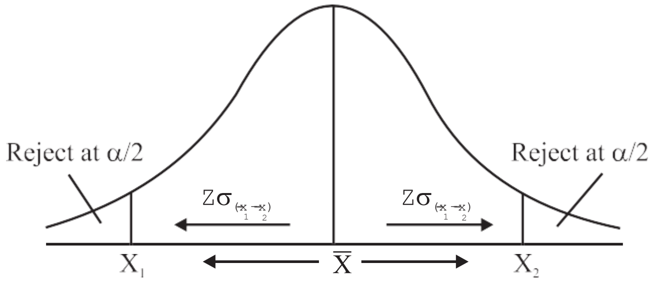
NOTES

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sigma_{\bar{X}_1 - \bar{X}_2}} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

or,
$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Then the decisions can be made on the basis of whether the value of Z so calculated falls within the region of acceptance or whether it falls in the region of rejection at a given value of the level of significance.

Another way to test for the significance of such a difference is to put one sample mean as the mean of the normal distribution and see if the second sample mean lies within the region of acceptance or not, at a given value of α. If the second sample mean lies within the acceptance region (within the bounds of X_1 and X_2 as shown below) then we can accept the null hypothesis that there is no significant difference between the two population means and that both samples come from the same population and any numerical difference in values of these two sample means happened by chance or due to a sampling error.



Example 8.12: A potential buyer of electric bulbs bought 100 bulbs each of two famous brands, A and B. Upon testing both these samples, he found that brand A had a mean life of 1500 hours with a standard deviation of 50 hours whereas brand B had an average life of 1530 hours with a standard deviation of 60 hours. Can it be concluded at 5% level of significance ($\alpha = 0.05$) that the two brands differ significantly in quality?

Solution: We assume that there is no significant difference in the quality of both brands so that brand A is as good as brand B in terms of average number of operating hours, so that,

Null hypothesis: $H_0 : \mu_1 = \mu_2$

Alternate hypothesis: $H_1 : \mu_1 \neq \mu_2$

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Where,
$$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

NOTES

$$= \sqrt{\frac{(50)^2}{100} + \frac{(60)^2}{100}}$$

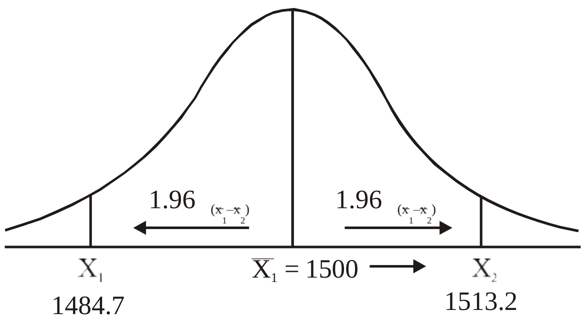
$$= \sqrt{61} = 7.81$$

Hence, $$Z = \frac{1500 - 1530}{7.81}$$

$$= -\left(\frac{30}{7.81}\right) = -(3.841)$$

Since the computed numerical absolute value of Z is more than the critical value of Z from the table at $\alpha = 0.05$, which is = 1.96, we cannot accept the null hypothesis.

We could also solve this problem by establishing the confidence interval where interval boundaries are as given:



In our case, we know that:

$$\sigma_{(\bar{X}_1 - \bar{X}_2)} = \qquad = 7.81$$

Then,

$$\text{Lower limit} = X_1 = \bar{X} - Z\sigma_{(\bar{X}_1 - \bar{X}_2)}$$

$$= 1500 - 1.96(7.81) = 1484.7$$

$$\text{Upper limit} = X_1 = \bar{X} - Z\sigma_{(\bar{X}_1 - \bar{X}_2)}$$

$$= 1500 + 1.96(7.81) = 1515.3$$

Since the value of $\bar{X}_2$ as 1530 lies beyond the acceptable limit of 1515.3, we cannot accept the null hypothesis.

Even though, our example above is a two-tailed test, we can also perform one-tailed tests for the differences between two population means where the null hypothesis is rejected when one mean is significantly higher or significantly lower than the other mean. These one-tailed tests are conceptually similar to the one-tailed tests of a single mean discussed earlier. We can illustrate this with the help of following example.

Example 8.13: A civil group in the city claims that a female college graduate earns less than a male college graduate. To test this claim, a survey of starting salary of 60 male graduates and 50 female graduates was taken and it was found that the average starting salary for female graduates was \$29,500 with a standard deviation of \$500 and the average salary for male graduates was \$30,000 with a standard deviation of \$600. At 1% level of significance, ($\alpha = 0.01$), test if the claim of this civil group is valid.

NOTES

Solution: The civil group claims that the average starting salary of female graduates is considerably less than the average starting salary of male graduates. The null hypothesis states that the starting salary of female graduates is not less than the starting salary of male graduates. Accordingly the null hypothesis will be rejected only if the average starting salary of female graduates is significantly less than the corresponding average starting salary of male graduates. The null hypothesis will not be rejected if this average is considerably higher than the average starting salary of male graduates. Hence, it is a one-tailed test.

Let $\bar{X}_1$ and s_1 represent the sample mean and the standard deviation, respectively, of the starting salary of female graduates.

Similarly, let $\bar{X}_2$ and s_2 respectively represent the mean and the standard deviation of the starting salary of male graduates. This data can be represented as follows:

Females	*Males*
$\bar{X}_1 = \$29,500$	$\bar{X}_2 = \$30,000$
$s_1 = \$500$	$s_2 = \$600$
$n_1 = 50$	$n_2 = 60$

We have to test whether at $\alpha = 0.01$, the observed difference between $\bar{X}_1$ and $\bar{X}_2$ is significant or not.

Hence,

$$H_0 : \mu_1 \geq \mu_2$$

$$H_1 : \mu_1 < \mu_2$$

Now,

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$$= \frac{29,500 - 30,000}{\sqrt{\frac{(500)^2}{50} + \frac{(600)^2}{60}}}$$

$$= -\left(\frac{500}{\sqrt{5000+6000}}\right)$$

$$= -\frac{500}{104.88} = -(4.77)$$

At 1% level of significance, the critical value of Z on the left-end of the tail for a one-tailed test is –(2.33). Since our computed absolute numerical value of Z is higher than the absolute critical value of Z, we cannot accept the null hypothesis. This is illustrated in the following Z score normal distribution curve.

NOTES

Testing for the Difference of Two Population Proportions

In some situations, it is necessary to check whether the two population proportions are equal or not. Suppose we want to check whether the percentage of female students entering college after completing high school is significantly different than the percentage of male students similarly entering college. Or suppose, that we want to test whether the proportion of people supporting a national political leader in the north of the country is similar to proportion of people supporting him in the south. These tests require comparisons of two proportions to see if any difference between them is significant or not.

Distribution of Differences in Proportions

Since we are trying to find out if the difference between two population proportions is significant or not, we need to know the distribution of *differences of sample proportions*, just as we did in the case of comparison of two sample means earlier. This concept can best be illustrated by an example.

Suppose that we select 10 random samples of 200 students each (n_1) at Medgar Evers college and record the proportion (p_1) of females in each sample. Similarly, we also select 10 samples of 200 students each (n_2) from Baruch College and record the proportion of females (p_2) for each sample. These proportions and their differences ($p_1 - p_2$) for each paired sample are tabulated below:

Sample	Medgar Evers (p_1)	Baruch (p_2)	Difference ($p_1 - p_2$)
1	0.64	0.57	0.07
2	0.65	0.60	0.05
3	0.58	0.58	0.00
4	0.62	0.65	–0.03
5	0.56	0.62	–0.06
6	0.66	0.61	0.05
7	0.60	0.55	0.05
8	0.59	0.59	0.00
9	0.62	0.57	0.05
10	0.58	0.56	0.02

The distribution of values of ($p_1 - p_2$) above is known as the distribution of *differences of sample proportions*. Theoretically, if we took all possible pairs of random samples from these two populations and found the proportion of females in these samples and calculated the differences in each sample ($p_1 - p_2$), then the resulting distribution of these differences will be approximately normally distributed with the following characteristics:

1. Since these proportions are represented by binomial distribution and we are approximating the binomial distribution to the normal distribution, the sample sizes from each population should be large enough. In general,

$$n_1p_1 \geq 5$$
$$n_2p_2 \geq 5$$
$$n_1q_1 \geq 5$$
$$n_2q_2 \geq 5$$

2. The mean of the distribution of differences of proportions is given by $(\pi_1 - \pi_2)$, where *px* equals the proportion of female students in the population of all students at Medgar Evers college and π_2 is the proportion of female students in the population of all students at Baruch college.

3. The standard deviation of the distribution of differences in proportions is denoted by $\hat{\sigma}_p$ (sigma sub *p* hat) and is given by:

$$=$$

Where (pi hat) is the pooled estimate of the values p_1 and p_2 under null hypothesis, which assumes that there is no difference between the two population proportions. This is given by,

$$=$$

$$\hat{\sigma}_p = \sqrt{\hat{\pi}(1-\hat{\pi})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$$

$$\hat{\pi} = \frac{n_1p_1 + n_2p_2}{n_1 + n_2}$$

Then using the test for Z scores, since it is approximated as normal distribution, we get,

$$Z = \frac{p_1 - p_2}{\hat{\sigma}_p}$$

We compare this computed value of Z with the critical value of Z from the table under a given level of significance and decide whether to accept or reject the null hypothesis.

Two-Tailed Test for Differences between Two Proportions

Example 8.14: A sample of 200 students at Baruch college revealed that 18% of them were seniors. A similar sample of 400 students at Hunter college revealed that 15% of them were seniors. To test whether the difference between these two proportions is significant enough to conclude that these populations are indeed different at 5% level of significance ($\alpha = 0.05$).

Solution: Null hypothesis: $H_0 : \pi_1 = \pi_2$

Alternate hypothesis: $H_1 : \pi_1 \neq \pi_2$

It is a two-tailed test because if the proportion of seniors at Baruch college is significantly higher than the proportion of seniors at Hunter college, then the null hypothesis will be rejected and similarly, if the proportion of seniors at Baruch college is significantly

NOTES

lower than the proportion of seniors at Hunter college, the null hypothesis will again be rejected.

Now,

The proportion of seniors at Baruch college = p_1 = 0.18

The proportion of seniors at Hunter college = p_2 = 0.15

Then,

$$Z = \frac{p_1 - p_2}{\hat{\sigma}_p}$$

Let us first calculate the value of $\hat{\sigma}_p$. We know that,

$$=$$

and,

$$=$$

Here,

n_1 = 200

n_2 = 400

p_1 = 0.18

p_2 = 0.15

Substituting these values, we get,

$$\hat{\pi} =$$

$$= \frac{36 + 60}{600} = \frac{96}{600} = 0.16$$

Substituting the value of $\hat{\pi}$ we calculate the value of as,

$$=$$

$$=$$

Now,

$$Z = \frac{p_1 - p_2}{\hat{\sigma}_p}$$

$$= \frac{0.18 - 0.15}{0.0317} = \frac{0.03}{0.0317} = 0.95$$

Since our computed value of Z is less than the critical value of Z = 1.96 at α = 0.05, for a two-tailed test, we cannot reject the null hypothesis.

One-Tailed Test for Difference between Two Proportions

Conceptually, the one-tailed test for differences between two population proportions is similar to a one-tailed test for the difference between two population means and the area of rejection will lie only in one end of the normal curve, either in the left end tail or in the right end tail, depending upon the type of problem.

Example 8.15: An insurance company believes that smokers have higher incidence of heart diseases than non-smokers in men over 50 years of age. Accordingly, it is considering to offer discounts on its life insurance policies to non-smokers. However, before the decision can be made, an analysis is undertaken to justify its claim that the smokers are at a higher risk of heart disease than non-smokers. The company randomly selected 200 men which included 80 smokers and 120 non-smokers. The survey indicated that 18 smokers suffered from heart disease and 15 non-smokers suffered from heart disease. At 5% level of significance, can we justify the claim of the insurance company that smokers have a higher incidence of heart disease than non-smokers?

Solution: Let p_1 be the proportion of male smokers over 50 years of age who suffer from heart disease in the entire population and let p_2 be the corresponding proportion of non-smokers. Then,

Null hypothesis: $H_0 : \pi_1 = \pi_2$

Alterante hypothesis: $H_1 : \pi_1 > \pi_2$

$$\text{Test statistic: } Z = \frac{p_1 - p_2}{\hat{\sigma}_p}$$

Now,

p_1 = Proportion of male smokers over 50 years of age who suffer from heart disease,

$$\sqrt{\hat{\pi}(1-\hat{\pi})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$$

$$\hat{\pi} = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$$

$$\frac{33}{200}$$

$$\hat{\pi} = 0.165$$

$$= \frac{18}{80} = 0.225$$

p_2 = Proportion of male non-smokers over 50 years of age who suffer from heart disease,

$$= \frac{15}{120} = 0.125$$

$$\hat{\sigma}_p =$$

$$=$$

$$= \frac{80(0.225) + 120(0.125)}{80 + 120}$$

$$= \frac{18 + 15}{200}$$

$$=$$

NOTES

Then, $\hat{\sigma}_p =$

$=$

$=$

Hence, $Z = \dfrac{0.225 - 0.125}{0.0536}$

$= \dfrac{0.1}{0.0536} = 1.86$

Since the critical value of Z at $\alpha = 0.05$ for a one-tailed test is 1.64 and since our computed value of $Z = 1.86$ is higher than the critical value of Z, we cannot accept the null hypothesis. It shows that there is a strong evidence to infer that the proportion of smokers who have heart diseases is greater than the proportion of non-smokers who have heart disease.

3.8 SUMMARY

3.8.1 One of the major objectives of statistical analysis is to know the 'true' values of different parameters of the population. Since it is not possible due to time, cost and other constraints to take the entire population for consideration, random samples are taken from the population.

3.8.2 Central Limit Theorem states that, 'Regardless of the shape of the population, the distribution of the sample means approaches the normal probability distribution as the sample size increases.'

3.8.3 Standard error of the mean is a measure of dispersion of the distribution of sample means and is similar to the standard deviation in a frequency distribution and it measures the likely deviation of a sample mean from the grand mean of the sampling distribution.

3.8.4 A hypothesis is an approximate assumption that a researcher wants to test for its logical or empirical consequences.

3.8.5 Hypothesis refers to a provisional idea whose merit needs evaluation, but has no specific meaning, though it is often referred as a convenient mathematical approach for simplifying cumbersome calculation.

3.8.6 Setting up and testing hypothesis is an integral art of statistical inference. Hypotheses are often statements about population parameters like variance and expected value.

3.8.7 According to Karl Popper, a hypothesis must be falsifiable and that a proposition or theory cannot be called scientific if it does not admit the possibility of being shown false.

3.8.8 Testing a statistical hypothesis on the basis of a sample enables us to decide whether the hypothesis should be accepted or rejected. The sample data enables us to accept or reject the hypothesis.

Check Your Progress

4. What is Chi-square test?
5. What conditions need to be fulfilled before using *t*-test?
6. Who developed *F*-test?
7. What are the two elements of *F* ratio?

NOTES

- A hypothesis stated in the hope of being rejected is called a null hypothesis and is denoted by H_0.
- If H_0 is rejected, it may lead to the acceptance of an alternative hypothesis denoted by H_1.
- In this Type I Error, you may reject a null hypothesis when it is true. It means rejection of a hypothesis, which should have been accepted.
- In this Type II Error, you are supposed to accept a null hypothesis when it is not true. It means accepting a hypothesis, which should have been rejected.
- Chi-square test is a non-parametric test of statistical significance for bivariate tabular analysis (also known as cross-breaks).
- Any appropriate test of statistical significance lets you know the degree of confidence you can have in accepting or rejecting a hypothesis.
- Sir William S. Gosset (pen name Student) developed a significance test and through it made significant contribution to the theory of sampling applicable in case of small samples. When population variance is not known, the test is commonly known as Student's *t*-test and is based on the *t* distribution.
- Unlike normal distribution, which is only one type of curve irrespective of the value of the mean and the standard deviation, the *F* distribution is a family of curves. A particular curve is determined by two parameters. These are the degrees of freedom in the numerator and the degrees of freedom in the denominator. The shape of the curve changes as the number of degrees of freedom changes.

3.9 KEY TERMS

3.9.1 **Sampling distribution:** A probability distribution of all possible sample means of a given size, selected from a population

3.9.2 **Standard error of mean:** Measures the likely deviation of a sample mean from the grand mean of the sampling distribution

3.9.3 **Null hypothesis:** A hypothesis stated in the hope of being rejected

3.9.4 **Hypothesis:** A statement or assumption concerning a population. For the purpose of decision-making, a hypothesis has to be verified and then accepted or rejected

3.9.5 **Type I Error:** In this type of error, you may reject a null hypothesis when it is true. It means rejection of a hypothesis, which should have been accepted

3.9.6 **Type II Error:** In this type of error, you are supposed to accept a null hypothesis when it is not true. It means accepting a hypothesis, which should have been rejected

3.9.7 **Chi-square test:** Any statistical hypothesis test in which the test statistics has a chi-square distribution when the null hypothesis is true

3.10 ANSWERS TO 'CHECK YOUR PROGRESS'

1. When it was possible to take all the possible samples of the same size, then the distribution of the results of these samples would be referred to as, 'sampling distribution'.

NOTES

2. Central Limit Theorem states that, 'Regardless of the shape of the population, the distribution of the sample means approaches the normal probability distribution as the sample size increases.'
3. A hypothesis is an approximate assumption that a researcher wants to test for its logical or empirical consequences. Hypothesis refers to a provisional idea whose merit needs evaluation, but has no specific meaning.
4. Chi-square test is a non-parametric test of statistical significance for bivariate tabular analysis (also known as cross-breaks).
5. The *t*-test is used when two conditions are fullfiled:
   (i) The sample size is less than 30, i.e., when $n \le 30$.
   (ii) The population standard deviation (σ_p) must be unknown.
6. *F*-test was developed by R.A. Fisher.
7. *F* ratio contains only two elements, which are the variance between the samples and the variance within the samples.

3.11 QUESTIONS AND EXERCISES

Short-Answer Questions

1. What do you understand by sampling distribution?
2. What are the characteristics of a hypothesis?
3. What is the test of significance?
4. Define the term 'degree of freedom'.
5. What are the areas of application of Chi-square test?
6. What are the important characteristics of Chi-square?
7. What are the two types of classifications involved in the analysis of variance?
8. What is *F*-distribution?
9. What are the one-tailed and two-tailed tests?

Long-Answer Questions

1. Discuss the properties of Central Limit Theorem.
2. Explain the two types of errors in statistical hypothesis.
3. Write an explanatory note on Chi-square test. When is it used?
4. What is *t*-test? What are the assumptions? Discuss.
5. Explain analysis of variance. Explain the two approaches to calculate the measure of variance.
6. Discuss the steps to calculate SSB.

3.12 FURTHER READING

Allen, R.G.D. 2008. *Mathematical Analysis For Economists*. London: Macmillan and Co., Limited.

Chiang, Alpha C. and Kevin Wainwright. 2005. *Fundamental Methods of Mathematical Economics*, 4 edition. New York: McGraw-Hill Higher Education.

Yamane, Taro. 2012. *Mathematics For Economists: An Elementary Survey*. USA: Literary Licensing.

Baumol, William J. 1977. *Economic Theory and Operations Analysis*, 4th revised edition. New Jersey: Prentice Hall.

Hadley, G. 1961. *Linear Algebra*, 1st edition. Boston: Addison Wesley.

Vatsa , B.S. and Suchi Vatsa. 2012. *Theory of Matrices*, 3rd edition. London: New Academic Science Ltd.

Madnani, B C and G M Mehta. 2007. *Mathematics for Economists*. New Delhi: Sultan Chand & Sons.

Henderson, R E and J M Quandt. 1958. *Microeconomic Theory: A Mathematical Approach*. New York: McGraw-Hill.

Nagar, A.L. and R.K.Das. 1997. *Basic Statistics*, 2nd edition. United Kingdom: Oxford University Press.

Gupta, S.C. 2014. *Fundamentals of Mathematical Statistics*. New Delhi: Sultan Chand & Sons.

M. K. Gupta, A. M. Gun and B. Dasgupta. 2008. *Fundamentals of Statistics*. West Bengal: World Press Pvt. Ltd.

Saxena, H.C. and J. N. Kapur. 1960. *Mathematical Statistics*, 1st edition. New Delhi: S. Chand Publishing.

Hogg, Robert V., Joeseph McKean and Allen T Craig. *Introduction to Mathematical Statistics*, 7th edition. New Jersey: Pearson.

NOTES

UNIT IV CORRELATION AND REGRESSION

Structure

4.0 INTRODUCTION

In this unit, you will learn about the correlation analysis techniques that analyses the indirect relationships in sample survey data and establishes the variables which are most closely associated with a given action or mindset. It is the process of finding how accurately the line fits using the observations. You will also learn about the scatter diagram, least squares method and standard error of the estimate. In this unit, you will also learn about regression analysis. Regression is a technique used to determine the statistical relationship between two (or more) variables and to make prediction of one variable on the basis of one or more other variables.

4.1 UNIT OBJECTIVES

After going through this unit, you will be able to:

- Understand what correlation is
- Explain different types of correlation

- Understand the different methods of studying correlation
- Describe correlation coefficient
- Evaluate coefficient of determination and coefficient of correlation
- Calculate correlation using various methods
- Understand regression analysis and its assumptions
- Describe the various regression models

4.2 CORRELATION

Correlation analysis is the statistical tool generally used to describe the degree to which one variable is related to another. The relationship, if any, is usually assumed to be a linear one. This analysis is used quite frequently in conjunction with regression analysis to measure how well the regression line explains the variations of the dependent variable. In fact, the word correlation refers to the relationship or interdependence between two variables. There are various phenomena which have relation to each other. When, for instance, demand of a certain commodity increases, then its price goes up and when its demand decreases then its price comes down. Similarly, with age the height of the children, with height the weight of the children, with money supply the general level of prices go up. Such sort of relationship can as well be noticed for several other phenomena. The theory by means of which quantitative connections between two sets of phenomena are determined is called the *Theory of Correlation.*

On the basis of the theory of correlation, one can study the comparative changes occurring in two related phenomena and their cause-effect relation can be examined. It should, however, be borne in mind that relationship like 'black cat causes bad luck', 'filled-up pitchers result in good fortune' and similar other beliefs of the people cannot be explained by the theory of correlation since they are all imaginary and are incapable of being justified mathematically. Thus, correlation is concerned with the relationship between two related and quantifiable variables. If two quantities vary in sympathy so that a movement (an increase or decrease) in the one tends to be accompanied by a movement in the same or opposite direction in the other and the greater the change in the one, the greater is the change in the other, the quantities are said to be correlated. This type of relationship is known as correlation or what is sometimes called, in statistics, as co-variation.

For correlation, it is essential that the two phenomena should have a cause-effect relationship. If such relationship does not exist then there can be no correlation. If, for example, the height of the students as well as the height of the trees increases, then one should not call it a case of correlation because the two phenomena, viz. the height of students and the height of trees are not even causally related. However, the relationship between the price of a commodity and its demand, the price of a commodity and its supply, the rate of interest and savings, etc., are examples of correlation since in all such cases the change in one phenomenon is explained by a change in other phenomenon.

It is appropriate here to mention that correlation in case of phenomena pertaining to natural sciences can be reduced to absolute mathematical terms, e.g., heat always increases with light. But in phenomena pertaining to social sciences it is often difficult to establish any absolute relationship between two phenomena. Hence, in social sciences we must take the fact of correlation being established if in a large number of cases, two variables always tend to move in the same or opposite direction.

Correlation can either be positive or it can be negative. Whether correlation is positive or negative would depend upon the direction in which the variables are moving. If both variables are changing in the same direction, then correlation is said to be positive but when the variations in the two variables take place in opposite direction, the correlation is termed as negative. This can be explained as follows:

Changes in Independent Variable	*Changes in Dependent Variable*	*Nature of Correlation*
Increase (+)↑	Increase (+)↑	Positive (+)
Decrease (–)↓	Decrease (–)↓	Positive (+)
Increase (+)↑	Decrease (–)↓	Negative (–)
Decrease (–)↓	Increase (+)↑	Negative (–)

Correlation can either be linear or it can be non-linear. The non-linear correlation is also known as curvilinear correlation. The distinction is based upon the constancy of the ratio of change between the variables. When the amount of change in one variable tends to bear a constant ratio to the amount of change in the other variable, then the correlation is said to be linear. In such a case if the values of the variables are plotted on a graph paper, then a straight line is obtained. This is why the correlation is known as linear correlation. But when the amount of change in one variable does not bear a constant ratio to the amount of change in the other variable, i.e., the ratio happens to be a variable instead of a constant, then the correlation is said to be non-linear or curvilinear. In such a situation, we shall obtain a curve if the values of the variables are plotted on a graph paper.

Correlation can either be simple correlation or it can be partial correlation or multiple correlation. The study of correlation for two variables (of which one is independent and the other is dependent) involves application of simple correlation. When more than two variables are involved in a study relating to correlation, then it can either be a multiple correlation or a partial correlation. Multiple correlation studies the relationship between a dependent variable and two or more independent variables. In partial correlation, we measure the correlation between a dependent variable and one particular independent variable assuming that all other independent variables remain constant.

Statisticians have developed two measures for describing the correlation between two variables, viz. the coefficient of determination and the coefficient of correlation.

4.3 DIFFERENT METHODS OF STUDYING CORRELATION

The following are the different methods of studying correlation and analysing the effects.

4.3.1 The Scatter Diagram

The scatter diagram is a graph of observed plotted points where each point represents the values of X and Y as a coordinate. It portrays the relationship between these two variables graphically. By looking at the scatter of the various points on the chart, it is possible to determine the extent of association between these two variables. The wider the scatter on the chart, the less close is the relationship. On the other hand, the closer the points and the closer they come to falling on a line passing through them, the higher the degree of relationship. If all the points fall on a line, the relationship is perfect. If this line goes up from the lower left-hand corner to the upper right-hand corner, i.e., if the slope of the line is

NOTES

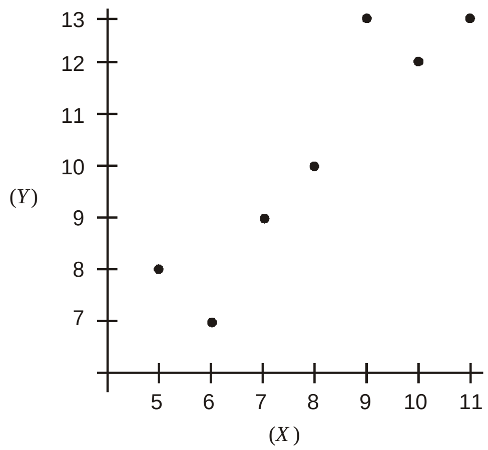
positive, then the correlation between the two variables is considered to be perfect positive. Similarly, if this line starts at the upper left-hand corner and comes down to the lower right-hand corner of the diagram, i.e., if the slope is negative, and also all points fall on the line, then their correlation is said to be perfect negative.

Example 9.1: The following data represents the money spent on advertising of a product and the respective profits realized from each advertising period for the given product. The amounts are in thousands of dollars. Assume profit to be a dependent variable and advertising as an independent variable.

Advertising (*X*)	Profit (*Y*)
5	8
6	7
7	9
8	10
91	13
10	12
11	13

Solution: We shall draw a scatter diagram for this data.

We can see that the trend in the relationship is increasing and even though this relationship is not perfect, i.e., all the points do not lie in a straight line, the profits in general do increase as the advertising budget increases. This gives us a reasonable visual idea about the relationship between *X* and *Y*.



4.3.2 The Linear Regression Equation

The pattern of the scatter diagram shown above indicates a linear relationship between *X* and *Y*, and this relationship can be described by a straight line through these points. This line is known as the *line of regression.* This line should be the most representative of the data. There are infinite number of lines that can approximately pass through this pattern, and we are looking for one line out of these, that is most suitable as representative of all the data. This line is known as the *line of best fit.* But, how do we find this regression line or the line of best fit? The best line would be the one that passes through all the points. Since that is not possible, we must find a line which is closest to all the points. A line will be closest to all these points if the total distance between the line and all the points is minimum. However, the same points will be above the line, so that the difference between the line and the points above the line would be positive and some points will be below the line, so that these differences would be negative. Accordingly, for the best line through this data, these differences will cancel each other, and the total sum of differences as a measure of best fit would not be valid. However, if we took these differences individually and squared them, this would eliminate the problem of positive and negative differences. Since the square of negative differences would also be positive, the total sum of squares would be positive.

Now, we are looking for a line which is closest to all the points. Hence, for such a line the absolute sum of differences between the points would be minimum and so would the sum of squares of these differences. Hence, this method of finding the line of best fit is known as the method of *least squares.*

This line of best fit is known as the regression line and the algebraic expression that identifies this line is a general straight line equation and is given as,

$$Y_c = b_0 + b_1X$$

where b_0 and b_1 are the two pieces of information called parameters which determine the position of the line completely. Parameter b_0 is known as the Y-intercept (or the value of Y_c at $X = 0$) and parameter b_1 determines the slope of the regression line which is the change in Y_c for each unit change in X.

Also, X represents a given value of the independent variable, and Y_c represents the computed value of the dependent variable based upon the above relationship.

This regression would have the following properties:

(a) $\Sigma (Y - Y_c) = 0$.
(b) $\Sigma (Y - Y_c)^2$ = Minimum.

where Y is the observed value of the dependent variable for a given value of X and Y_c is the computed value of the dependent variable for the same value of X. This relation between Y and Y_c is shown in Figure 9.1.

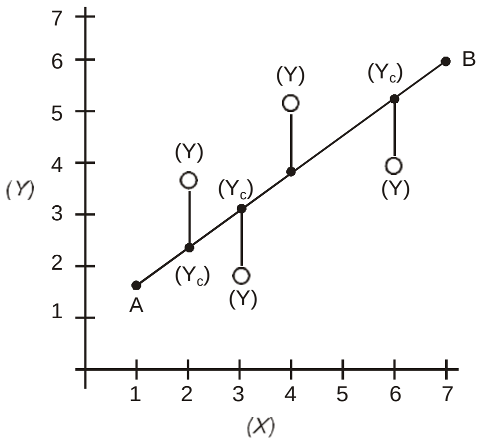


Fig. 9.1 *Observed and Computed Value of Dependent Variable*

The line AB is the line of best fit when,

(a) $\Sigma (Y - Y_c) = 0$.
(b) $\Sigma (Y - Y_c)^2$ = Minimum.

Here, Y is the actual observation and Y_c is the corresponding computed value, based upon the method of least squares.

Now, since $Y_c = b_0 + b_1X$ is the algebraic equation for any line, we must find the unique values of b_0 and b_1, which would automatically give us the *regression* line. These

NOTES

unique values of b_0 and b_1 based upon the *least squares* principle are calculated according to the following formulae:

$$b_0 = \frac{(\ Y)(\ X^2) \quad (\ X)(\ XY)}{n(\ X^2) \quad (\ X)^2}$$

and

$$b_1 = \frac{n(\ XY) \quad (\ X)(\ Y)}{n(\ X^2) \quad (\ X)^2}$$

The value of b_0 can also be calculated easily, once the value of b_1 has been calculated as follows:

$$b_0 = \overline{Y} \quad b_1\overline{X}$$

where $\overline{Y}$ and are simple arithmetic means of the Y data and X data respectively, and n represents the number of paired observations.

We can illustrate these calculations by an example.

Example 9.2: A researcher wants to find out if there is a relationship between the heights of the sons and the heights of their fathers. In other words, do tall fathers have tall sons? He took a random sample of 6 fathers and their 6 sons. Their heights in inches are given in an ordered array as follows.

Father (X)	Son (Y)
63	66
65	68
66	65
67	67
67	69
68	70

(a) For this data, compute the regression line.

(b) Based upon the relationship between the heights, what would be the estimate of the height of the son, if the father's height is 70 inches?

Solution: (a) We can start with showing the scatter diagram for this data.

The scatter diagram shows an increasing trend through which the line of the best fit *AB* can be established. This line is identified by:

$$Y_c = b_0 + b_1X$$

Where, $b_1 =$

And, $b_0 = \overline{Y} \quad b_1\overline{X}$

Let us make a table to calculate all these values.

X	Y	X^2	XY	Y^2
63	66	3969	4158	4356
65	68	4225	4420	4624
66	65	4356	4290	4225
67	67	4489	4489	4489
67	69	4489	4623	4761
68	70	4624	4760	4900
$\Sigma X = 396$	$\Sigma Y = 405$	$\Sigma X^2 = 26152$	$\Sigma XY = 26740$	$\Sigma Y^2 = 27355$

Then,

$$b_1 = \frac{6(26740) \quad (396)(405)}{6(26152) \quad (396)(396)}$$

$$= \frac{160440 \quad 160380}{156912 \quad 156816} \qquad = \qquad = 0.625$$

and, $b_0 =$

$\frac{405}{96}$ $0.625(396/6)$

$$= 62.5 - 41.25 \qquad = 26.25$$

Hence, the line of regression equation would be:

$$Y_c = b_0 + b_1X$$
$$= 26.25 + 0.625X$$

(b) If the father's height is 70 inches, i.e., if $X = 70$, then the computed height of the son or Y_c would be:

$$Y_c = 26.25 + 0.625\,(70)$$
$$= 26.25 + 43.75 = 70$$

Standard Error of the Estimate

We have found a line through the scatter points which best fits the data. But how good is this fit? How reliable is the estimated value of Y_c? How close are the values of Y_c to the observed values of Y? The closer these values are to each other, the better the fit. This means that if the points in the scatter diagram are closely spaced around the regression line, then the estimated value Y_c will be close to the observed value of Y and hence, this estimate can be considered as highly reliable. Accordingly, a measure of variability of scatter around the regression line would determine the reliability of this estimate Y_c. The smaller this estimate, the more dependable the prediction will be. This measure is similar in nature to standard deviation which is also a measure of scattered data around the mean.

This measure is known as *standard error of the estimate* and is used to determine the dispersion of observed values of Y about the regression line. This measure is designated by $S_{y.x}$ and is given by:

$$S_{y.x} = \sqrt{\frac{(Y \quad Y_c)^2}{n \quad 2}}$$

NOTES

Where

Y = Observed value of the dependent variable.

Y_c = Corresponding computed value of the dependent variable.

n = Sample size.

And,

$(n-2)$ = Degrees of freedom.

Based upon this relationship, a simpler formula for calculating $S_{y.x}$ would be:

$$S_{y.x} = \sqrt{\frac{(Y)^2 \quad b_0(\ Y) \quad b_1(\ XY)}{n \quad 2}}$$

Example 9.3: Considering Example 9.2, regarding the relationship of heights between sons and their fathers, calculate the standard error of the estimate $S_{y.x}$.

Solution: Now,

$$S_{y.x} = \sqrt{\frac{(Y)^2 \quad b_0(\ Y) \quad b_1(\ XY)}{n \quad 2}}$$

$$= \sqrt{\frac{27355 \quad 26.25(405) \quad 0.625(26740)}{4}}$$

$$= \sqrt{\frac{11.25}{4}} \qquad = \qquad = 1.678$$

4.4 CORRELATION COEFFICIENT

The coefficient of correlation symbolically denoted by '*r*' is another important measure to describe how well one variable is explained by another. It measures the degree of relationship between the two causally-related variables. The value of this coefficient can never be more than +1 or less than –1. Thus +1 and –1 are the limits of this coefficient. For a unit change in independent variable, if there happens to be a constant change in the dependent variable in the same direction, then the value of the coefficient will be +1 indicative of the perfect positive correlation; but if such a change occurs in the opposite direction, the value of the coefficient will be –1, indicating the perfect negative correlation. In practical life the possibility of obtaining either a perfect positive or perfect negative correlation is very remote particularly in respect of phenomena concerning social sciences. If the coefficient of correlation has a zero value then it means that there exists no correlation between the variables under study.

There are several methods of finding the coefficient of correlation but the following ones are considered important:

(i) Coefficient of Correlation by the Method of Least Squares.

(ii) Coefficient of Correlation using Simple Regression Coefficients.

(iii) Coefficient of Correlation through Product Moment Method or Karl Pearson's Coefficient of Correlation.

Whichever of these above-mentioned three methods we adopt, we get the same value of *r*.

4.4.1 Coefficient of Correlation by the Method of Least Squares

Under this method, first of all the estimating equation is obtained using the least square method of simple regression analysis. The equation is worked out as:

Total variation $Y \quad \bar{Y} \quad ^2$

Unexplained variation $Y \quad \hat{Y} \quad ^2$

Explained variation $\hat{Y} \quad \bar{Y} \quad ^2$

Then, by applying the following formulae we can find the value of the coefficient of correlation.

$$r = \sqrt{r^2} \quad \sqrt{\frac{\text{Explained variation}}{\text{Total variation}}}$$

$$= \sqrt{1 \quad \frac{\text{Unexplained variation}}{\text{Total variation}}}$$

$$= \sqrt{1 \quad \frac{Y \quad \hat{Y} \quad ^2}{Y \quad \bar{Y} \quad ^2}}$$

This clearly shows that coefficient of correlation happens to be the squareroot of the coefficient of determination.

Short-cut formula for finding the value of 'r' by the method of least squares can be repeated and readily written as follows:

$\hat{Y} \quad a \quad bX_i$

$$r = \sqrt{\frac{a \quad Y \quad b \quad XY \quad n\bar{Y}^2}{Y^2 \quad n\bar{Y}^2}}$$

Where
a = Y-intercept
b = Slope of the estimating equation
X = Values of the independent variable
Y = Values of dependent variable
$\bar{Y}$ = Mean of the observed values of Y
n = Number of items in the sample (i.e., pairs of observed data)

The plus (+) or the minus (–) sign of the coefficient of correlation worked out by the method of least squares is related to the sign of 'b' in the estimating equation, viz., $\hat{Y} \quad a \quad bX_i$. If '$b$' has a minus sign, the sign of 'r' will also be minus but if 'b' has a plus sign, then the sign of 'r' will also be plus. The value of 'r' indicates the degree along with the direction of the relationship between the two variables X and Y.

4.4.2 Coefficient of Correlation using Simple Regression Coefficient

Under this method, the estimating equation of Y and the estimating equation of X is worked out using the method of least squares. From these estimating equations we find the regression coefficient of X on Y, i.e., the slope of the estimating equation of X (symbolically written as b_{XY}) and this is equal to $r\frac{X}{Y}$ and similarly, we find the regression coefficient of Y on X, i.e., the slope of the estimating equation of Y (symbolically written as b_{YX}) and this is equal to . For finding 'r', the square root of the product of these two regression coefficients are worked out as stated below:[1]

$$r = \qquad =$$

$$= \qquad = r$$

As stated earlier, the sign of 'r' will depend upon the sign of the regression coefficients. If they have minus sign, then 'r' will take minus sign but the sign of 'r' will be plus if regression coefficients have plus sign.

4.4.3 Karl Pearson's Coefficient of Correlation

Karl Pearson's method is the most widely-used method of measuring the relationship between two variables. This coefficient is based on the following assumptions:

- There is a linear relationship between the two variables which means that straight line would be obtained if the observed data are plotted on a graph.
- The two variables are causally related which means that one of the variables is independent and the other one is dependent.
- A large number of independent causes are operating in both the variables so as to produce a normal distribution.

According to Karl Pearson, 'r' can be worked out as under:

$$r =$$

Where $X = (X - \bar{X})$

$Y = (Y - \bar{Y})$

σ_X = Standard deviation of X series and is equal to $\sqrt{\frac{X^2}{n}}$

σ_Y = Standard deviation of Y series and is equal to $\sqrt{\frac{\sum Y^2}{n}}$

n = Number of pairs of X and Y observed.

1. Remember the short-cut formulae to workout b_{XY} and b_{YX}:

$$b_{XY} = \frac{\sum XY - n\bar{X}\bar{Y}}{\sum Y^2 - n\bar{Y}^2}$$

and $$b_{YX} = \frac{\sum XY - n\bar{X}\bar{Y}}{\sum X^2 - n\bar{X}^2}$$

A short-cut formula known as the Product Moment Formula (PMF) can be derived from the above-stated formula as under:

$$r = \frac{\sum XY}{n\sigma_X \sigma_Y} =$$

$$= \frac{\sum XY}{\sqrt{\sum X^2 \sum Y^2}}$$

The above formulae are based on obtaining true means (viz., $\bar{X}$ and $\bar{Y}$) first and then doing all other calculations. This happens to be a tedious task particularly if the true means are in fractions. To avoid difficult calculations, we make use of the assumed means in taking out deviations and doing the related calculations. In such a situation, we can use the following formula for finding the value of '*r*':

(a) In Case of Ungrouped Data:

$$r = \frac{\frac{\sum dX.dY}{n} - \left(\frac{\sum dX}{n} \cdot \frac{\sum dY}{n}\right)}{\sqrt{\frac{\sum dX^2}{n} - \left(\frac{\sum dX}{n}\right)^2} \cdot \sqrt{\frac{\sum dY^2}{n} - \left(\frac{\sum dY}{n}\right)^2}}$$

$$= \frac{\sum dX.dY - \frac{\sum dX \cdot \sum dY}{n}}{\sqrt{\sum dX^2 - \frac{(\sum dX)^2}{n}} \cdot \sqrt{\sum dY^2 - \frac{(\sum dY)^2}{n}}}$$

Where $\sum dX = \sum(X - X_A)$ $\quad X_A$ = Assumed average of X

$\sum dY = \sum(Y - Y_A)$ $\quad Y_A$ = Assumed average of Y

$\sum dX^2 = \sum(X - X_A)^2$

$\sum dY^2 = \sum(Y - Y_A)^2$

$\sum dX \cdot dY = \sum(X - X_A)(Y - Y_A)$

n = Number of pairs of observations of X and Y

(b) In Case of Grouped Data:

$$r = \frac{\frac{\sum fdX.dY}{n} - \left(\frac{\sum fdX}{n} \cdot \frac{\sum fdY}{n}\right)}{\sqrt{\frac{\sum fdX^2}{n} - \left(\frac{\sum fdX}{n}\right)^2} \sqrt{\frac{\sum fdY^2}{n} - \left(\frac{\sum fdY}{n}\right)^2}}$$

Or

$$r = \frac{\sum fdX.dY - \frac{\sum fdX \cdot \sum fdY}{n}}{\sqrt{\sum fdX^2 - \frac{(\sum fdX)^2}{n}} \cdot \sqrt{\sum fdY^2 - \frac{(\sum fdY)^2}{n}}}$$

Where, $\sum fdX.dY = \sum f(X - X_A)(Y - Y_A)$

$\sum fdX = \sum f(X - X_A)$

$\sum fdY = \sum f(Y - Y_A)$

$\sum fdY^2 = \sum f(Y - Y_A)^2$ $\quad \sum fdX^2 = \sum f(X - X_A)^2$

n = Number of pairs of observations of X and Y

NOTES

4.4.4 Probable Error of the Coefficient of Correlation

Probable Error (PE) of r is very useful in interpreting the value of r and is worked out as under for Karl Pearson's coefficient of correlation:

If r is less than its PE, it is not at all significant. If r is more than PE, there is correlation. *If r is more than 6 times its PE and greater than ± 0.5, then it is considered significant.*

Example 9.4: From the following data calculate 'r' between X and Y applying the following three methods:

4.4.4.1 The method of least squares.

4.4.4.2 The method based on regression coefficients.

4.4.4.3 The product moment method of Karl Pearson.

Verify the obtained result of any one method with that of another.

X	1	2	3	4	5	6	7	8	9
Y	9	8	10	12	11	13	14	16	15

Solution: Let us develop the following table for calculating the value of 'r':

X	Y	X^2	Y^2	XY
1	9	1	81	9
2	8	4	64	16
3	10	9	100	30
4	12	16	144	48
5	11	25	121	55
6	13	36	169	78
7	14	49	196	98
8	16	64	256	128
9	15	81	225	135
$n = 9$				
$\sum X = 45$	$\sum Y = 108$	$\sum X^2 = 285$	$\sum Y^2 = 1356$	$\sum XY = 597$

$\therefore \quad \bar{X} = 5; \qquad \bar{Y} = 12$

4.4.4.3.1 *Coefficient of correlation by the method of least squares is worked out as under*:

First of all find out the estimating equation

$$\hat{Y} = a + bX_i$$

Where, $b =$

$$= \frac{597 \quad 9\ 5 \quad 12}{285 \quad 9\ 25} \quad \frac{597 \quad 540}{285 \quad 225} = \frac{57}{60} = 0.95$$

And, $a = \bar{Y} - b\bar{X}$

$$= 12 - 0.95(5) = 12 - 4.75 = 7.25$$

Hence, $$\hat{Y} = 7.25 + 0.95X_i$$

NOTES

Now, 'r' can be worked out as under by the method of least squares:

$$r =$$

$$= \sqrt{1 \quad \frac{Y \quad \hat{Y}^{\,2}}{Y \quad \bar{Y}^{\,2}}} \quad = \sqrt{\frac{\hat{Y} \quad \bar{Y}^{\,2}}{Y \quad \bar{Y}^{\,2}}}$$

$$= \sqrt{\frac{a \quad Y \quad b \quad XY \quad n\bar{Y}^2}{Y^2 \quad n\bar{Y}^2}}$$

As per the short-cut formula,

$$r = \sqrt{\frac{7.25(108) + 0.95(597) - 9(12)^2}{1356 - 9(12)^2}}$$

$$= \sqrt{\frac{783 \quad 567.15 \quad 1296}{1356 \quad 1296}} \qquad = \sqrt{\frac{54.15}{60}}$$

$$= \qquad = 0.95$$

4.4.4.3.2 *Coefficient of correlation by the method based on regression coefficients isworked out as under*:

Regression coefficients of Y on X:

$$b_{YX} =$$

$$= \frac{597 \quad 9 \quad 5 \quad 12}{285 \quad 9 \quad 5} \quad \frac{597 \quad 540}{285 \quad 225} \quad \frac{57}{60}$$

$\sqrt{1-\times}$ 0.9025 Explained variation $= \quad = 0.95$ 597 57 X 57 60 2 60 Total variation 60 $X^2 \quad n\bar{X}^2$

Regression coefficient of X on Y:

$$b_{XY} = \frac{XY \quad n\overline{XY}}{Y^2 \quad n\bar{Y}^2}$$

$$= \frac{597 \quad 9 \quad 5 \quad 12}{1356 \quad 9 \quad 12^{\,2}} \quad \frac{597 \quad 540}{1356 \quad 1296} \quad \frac{57}{60}$$

Hence, $\qquad r = \sqrt{b_{YX} . b_{XY}} \qquad =$

4.4.4.3.3 Coefficient of correlation by the product moment method of Karl Pearson isworked out as under:

$$r = \frac{XY \quad n\overline{XY}}{\sqrt{X^2 \quad n\bar{X}^2}\sqrt{Y^2 \quad n\bar{Y}^2}}$$

$$= \frac{597 \quad 9 \quad 5 \quad 12}{\sqrt{285 \quad 9 \quad 5^{\,2}}\sqrt{1356 \quad 9 \quad 12^{\,2}}}$$

$$= \frac{597 \quad 540}{\sqrt{285 \quad 225}\sqrt{1356 \quad 1296}} \quad \frac{57}{\sqrt{60}\sqrt{60}} = \frac{57}{60} = 0.95$$

Hence, we get the value of $r = 0.95$. We get the same value applying the other two methods also. Therefore, whichever method we apply, the results will be the same.

NOTES

Example 9.5: Calculate the coefficient of correlation and lines of regression from the following data.

Y *Sales Revenue* (Rs '000)	*X* *Advertising Expenditure* (*Rs '00*) 5–15	15–25	25–35	35–45	Total
75–125	3	4	4	8	19
125–175	8	6	5	7	26
175–225	2	2	3	4	11
225–275	2	3	2	2	9
Total	15	15	14	21	$n = 65$

Solution: Since the given information is a case of bivariate grouped data we shall extend the given table rightwards and downwards to obtain various values for finding '*r*' as stated below:

Y Sales Revenue (*Rs '000*)	*X* *Advertising Expenditure* (Rs'00) *5-15*	*15-25*	*25-35*	*35-40*	*Total* (*f*)	*Midpoint* of *Y*	*If* $A = 200$ $i = 50$ $\therefore dY$	*fdY*	fdY^2	*fdX.dY*
75-125	3	4	4	8	19	100	–2	–38	76	4
125-175	8	6	5	7	26	150	–1	–26	26	15
175-225	2	2	3	4	11	200	0	0	0	0
225-275	2	3	2	2	9	250	1	9	9	–5
Total (or *f*)	15	15	14	21	$n = 65$			$\sum fdY = -55$	$\sum fdY^2 = 111$	$\sum fdX.dY = 14$
Midpoint of *X*	10	20	30	40						
If $A = 30$ $i = 10$ $\therefore dX$	–2	–1	0	1						
fdX	–30	–15	0	21	$\sum fdX = -24$					
fdX^2	60	15	0	21	$\sum fdX^2 = 96$					
fdX.dY	24*	11	0	–21	$\sum fdX \cdot dY = 14$					

* This value has been worked out as under:

f	*dX.dY*	=	*f .dX dY*
(3)	(–2) (–2)	=	12
(8)	(–2) (–1)	=	16
(2)	(–2) (0)	=	0
(2)	(–2) (1)	=	–4
	Total		24

Similarly, for other columns also, the *f.dXdY* values can be obtained. The process can be repeated for finding *f.dX.dY* values row-wise and finally Σ*fdXdY* can be checked.

NOTES

$$\therefore \quad r = \frac{\dfrac{\sum fdX.dY}{n} - \left(\dfrac{\sum fdX}{n}\dfrac{\sum fdY}{n}\right)}{\sqrt{\dfrac{\sum fdX^2}{n} - \left(\dfrac{\sum fdX}{n}\right)^2}\sqrt{\dfrac{\sum fdY^2}{n} - \left(\dfrac{\sum fdY}{n}\right)^2}}$$

Putting the calculated values in the above equation we have:

$$r = \frac{0.2134 - \left(\dfrac{-24}{65}\times\dfrac{-55}{65}\right)}{\sqrt{\dfrac{96}{65} - \left(\dfrac{-24}{65}\right)^2}\sqrt{\dfrac{111}{65} - \left(\dfrac{-55}{65}\right)^2}}$$

$$= \frac{0.2134 - 0.3124}{\sqrt{1.48 - 0.14}\sqrt{1.71 - 0.72}}$$

$$= \frac{0.0970}{\sqrt{1.34 \quad 99}} \quad \frac{0.00970}{\sqrt{1.3266}} \quad \frac{0.0970}{1.15} \quad 0.0843$$

Hence, $r = (–)0.0843$

This shows a poor negative correlation between the two variables. Since only 0.64% (r^2) = $(0.08)^2 = 0.0064$) variation in Y (Sales revenue) is explained by variation in X (Advertising expenditure).

The two lines of regression are as under:

Regression line of X on Y: $X - \overline{X} = r\dfrac{\sigma_X}{\sigma_Y}(Y - \overline{Y})$

Regression line of Y on X: $Y - \overline{Y} = r\dfrac{\sigma_Y}{\sigma_X}(X - \overline{X})$

First obtain the following values:

$$\overline{X} = A + \frac{\sum fdX}{n}.i = 30 + \frac{(-24)}{65}\times 10 = 26.30$$

$$\overline{Y} = A + \frac{\sum fdY}{n}.i = 200 + \frac{-55}{65}\times 10 = 157.70$$

$$\sigma_X = \sqrt{\frac{\sum fdX^2}{n} - \left(\frac{\sum fdX}{n}\right)^2}\times i = \sqrt{\frac{96}{65} - \left(\frac{-24}{65}\right)^2}\times 10 = 11.60$$

$$\sigma_Y = \sqrt{\frac{\sum fdY^2}{n} - \left(\frac{\sum fdY}{n}\right)^2}\times i = \sqrt{\frac{111}{65} - \left(\frac{-55}{65}\right)^2}\times 50 = 49.50$$

Therefore, the regression line of X on Y:

$$(X - 26.30) = \frac{11.6}{49.5}(-0.084)(Y - 157.70)$$

Or, $\hat{X} = -0.02Y + 3.15 + 26.30$

$\therefore$ $\hat{X} = -0.02Y + 29.45$

Regression line of Y on X:

$$Y - 157.70 = \frac{49.5}{11.6}(-0.084)(X - 26.30)$$

Or, $\hat{Y} = -0.36X + 9.47 + 157.70$

$\therefore$ $\hat{Y} = -0.36X + 167.17$

NOTES

4.5 SPEARMAN'S RANK CORRELATION COEFFICIENT

If observations on two variables are given in the form of ranks and not numerical values, it is possible to compute what is known as *rank correlation* between the two series.

The rank correlation, written as ρ, is a descriptive index of agreement between ranks over individuals. It is the same as the ordinary coefficient of correlation computed on ranks, but its formula is simpler.

$$\rho = 1 - \frac{6\Sigma D_i^2}{n(n^2 - 1)}$$

where n is the number of observations and D_i the positive difference between ranks associated with the individuals i.

Like r, the rank correlation lies between –1 and +1.

Example 9.6: The ranks given by two judges to 10 individuals are as follows:

Individual	*Rank given by Judge I* x	*Judge II* y	$D = x - y$	D^2
1	1	7	6	36
2	2	5	3	9
3	7	8	1	1
4	9	10	1	1
5	8	9	1	1
6	6	4	2	4
7	4	1	3	9
8	3	6	3	9
9	10	3	7	49
10	5	2	3	9
				$\Sigma D^2 = 128$

Solution: The Rank Correlation is given by,

$$\rho = 1 - \frac{6\Sigma D^2}{n^3 - n} = 1 - \frac{6 \times 128}{10^3 - 10} = 1 - 0.776 = 0.224$$

The value of ρ = 0.224 shows that the agreement between the judges is not high.

Example 9.7: Referring to the previous case, compute r and compare.

Solution: The simple coefficient of correlation r for the previous data is calculated as follows:

x	y	x^2	y^2	xy
1	7	1	49	7
2	5	4	25	10
7	8	49	64	56
9	10	81	100	90
8	9	64	81	72
6	4	36	16	24
4	1	16	1	4
3	6	9	36	18
10	3	100	9	30
5	2	25	4	10
$\Sigma x = 55$	$\Sigma y = 55$	$\Sigma x^2 = 385$	$\Sigma y^2 = 385$	$\Sigma xy = 321$

$$r = \frac{321 - 10 \times \frac{55}{10} \times \frac{55}{10}}{\sqrt{385 - 10 \times \left(\frac{55}{10}\right)^2}\sqrt{385 - 10 \times \left(\frac{55}{10}\right)^2}} = \frac{18.5}{\sqrt{82.5 \times 82.5}} = \frac{18.5}{82.5} = 0.224$$

This shows that the Spearman ρ for any two sets of ranks is the same as the Pearson *r* for the set of ranks. But it is much easier to compute ρ.

Often, the ranks are not given. Instead, the numerical values of observations are given. In such a case, we must attach the ranks to these values to calculate ρ.

Example 9.8: On the basis of the table given below, analyse the type of correlation and calculate the group of equal ranks.

Marks in Maths	*Marks in Stats*	*Rank in Maths*	*Rank in Stats*	*D*	D^2
45	60	4	2	2	4
47	61	3	1	2	4
60	58	1	3	2	4
38	48	5	4	1	1
50	46	2	5	3	9
				$\Sigma D^2 = 22$	

Solution: The correlation can be analysed as,

$$\rho = 1 - \frac{6\Sigma D^2}{m^3 - m} = 1 - \frac{6 \times 22}{125 - 5} = -0.05$$

This shows a negative, though small, correlation between the ranks.

If two or more observations have the same value, their ranks are equal and obtained by calculating the means of the various ranks.

If in this data, marks in maths is 45 for each of the first two students, the rank of each would be $\frac{3+4}{2} = 3.5$. Similarly, if the marks of each of the last two students in statistics is 48, their ranks would be $\frac{4+5}{2} = 4.5$.

The problem takes the following shape:

Marks in Maths	*Marks in Stats*	*Rank* *x*	*y*	*D*	D^2
45	60	3.5	2	1.5	2.25
45	61	3.5	1	2.5	6.25
60	58	1	3	2	4.00
38	48	5	4.5	1.5	2.25
50	48	2	4.5	2.5	6.25

The formula which can be used in cases of equal ranks is,

$$\rho = 1 - \frac{6}{n^3 - n}\left[\Sigma D^2 + \frac{1}{12}\Sigma(m^3 - m)\right]$$

where $\frac{1}{12}\Sigma(m^3 - m)$ is to be added to ΣD^2 for each group of equal ranks, m being the number of equal ranks each time.

NOTES

For the given data, we have for x series, number of equal ranks $m = 2$

For y series also, $m = 2$, so that,

ρ =

= = = -0.1

4.6 CONCURRENT DEVIATION METHOD

When deviation is noted in two series representing some event, or some variation over some time interval, some deviation is found. Concurrent deviation is not concerned with the quantity of deviation; rather it is concerned with the direction of deviation. Take for example; we measure the thickness of a plate. We take different readings and note these in a tabular form. We find some variation in the observed reading every time. This is deviation. Now if we give the same thing to another person doing the same thing and both do it at the same time.

Concurrent deviation is noted by combining deviations in both the cases. Each subsequent reading are noted whether it is more than, equal to or less than its previous reading. If more, it is positive deviation and if less, then it is a negative deviation. In case there is no change in value there is no deviation. Here, no quantity is involved and only direction in which changes occur is noted. Combining both, we get concurrent deviation and a property called correlation coefficient for concurrent deviation is found.

This method of concurrent deviation is very easy and simple method to calculate coefficient of concurrent deviation, which finds application in business and commerce. In this method, correlation is calculated to find the direction of deviation and not the magnitude. If deviation of two time series is concurrent, their curves would move in the same direction and it indicates a positive correlation between them. Coefficient of concurrent deviations is calculated on this principle and its ordinality. It shows the relationship between short time fluctuations only.

The method involves following steps:

1. Deviation of both the series is calculated separately. Deviation of every item of a series depends on the value of previous item. If second item is higher than the first, then, it is shown by placing '+' sign against the second item in a new column with a header as **deviation *dx***. If smaller the '-' sign is to be put and if equal then '=' sign which means no change. The process is continued for the whole series.
2. Compute deviation in second series and show them in another column under the header by **deviation *dy***.
3. Construct another column for product of dx and dy $(dx.dy)$. This column denotes concurrent deviation.
4. Find number of pairs of concurrent deviations (C).

Check Your Progress

1. List the different types of correlations.
2. What is a scatter diagram method?
3. What do you mean by coefficient of correlation?
4. What is rank correlation coefficient?

NOTES

5. Use the following formulae:

$$r_c = \pm\sqrt{\pm\frac{2C-N}{N}}$$

Where, r_c = Coefficient of concurrent deviation.
C = Number of concurrent deviation.
and
N = Number of pairs of deviation.
Use of '+' and negative sign depends on the sign of

$\left(\frac{2C-N}{N}\right)$ which lies between –1 and +1.

Example 9.9: Find coefficient of correlation from the following data by the concurrent deviation method:

X	85	91	56	72	95	76	89	51	59	90
Y	18.3	20.8	16.9	15.7	19.2	18.1	17.5	14.9	18.9	15.4

Solution: Make a table of *X*, *Y*, *dx*, *dy* and *dxdy* as below:

X	Deviation *X* Series (*dx*)	*Y*	Deviation *Y* Series (*dy*)	Concurrent Deviation (*dxdy*)
85		18.3		
91	+	20.8	+	+
56	–	16.9	–	+
72	+	15.7	–	–
95	+	19.2	+	+
76	–	18.1	–	+
89	+	17.5	–	–
51	–	14.9	–	+
59	+	18.9	+	+
90	+	15.4		
	N = 9			C = 6

Putting values in the formulae, we get r_c as below:

$$r_c = +\sqrt{+\left(\frac{2\times6-9}{9}\right)}$$

$$r_c = +\sqrt{+\frac{3}{9}} = 0.577$$

4.7 COEFFICIENT OF DETERMINATION

The coefficient of determination (symbolically indicated as r^2, though some people would prefer to put it as R^2) is a measure of the degree of linear association or correlation between two variables, say *X* and *Y*, one of which happens to be an independent variable

NOTES

and the other a dependent variable. This coefficient is based on the following two kinds of variations:

(a) The variation of the Y values around the fitted regression line, viz. $\Sigma\left(Y-\hat{Y}\right)^2$, is technically known as the unexplained variation.

(b) The variation of the Y values around their own mean, viz. $\Sigma\left(Y-\overline{Y}\right)^2$, is technically known as the total variation.

If we subtract the unexplained variation from the total variation, we obtain what is known as the explained variation, i.e., the variation explained by the line of regression. Thus, Explained Variation = (Total variation) – (Unexplained variation)

$$=\Sigma\left(Y-\overline{Y}\right)^2-\Sigma\left(Y-\hat{Y}\right)^2$$

$$\hat{Y} \quad \overline{Y}^{2}$$

The Total and Explained as well as Unexplained variations can be shown as given in the Figure 9.2.

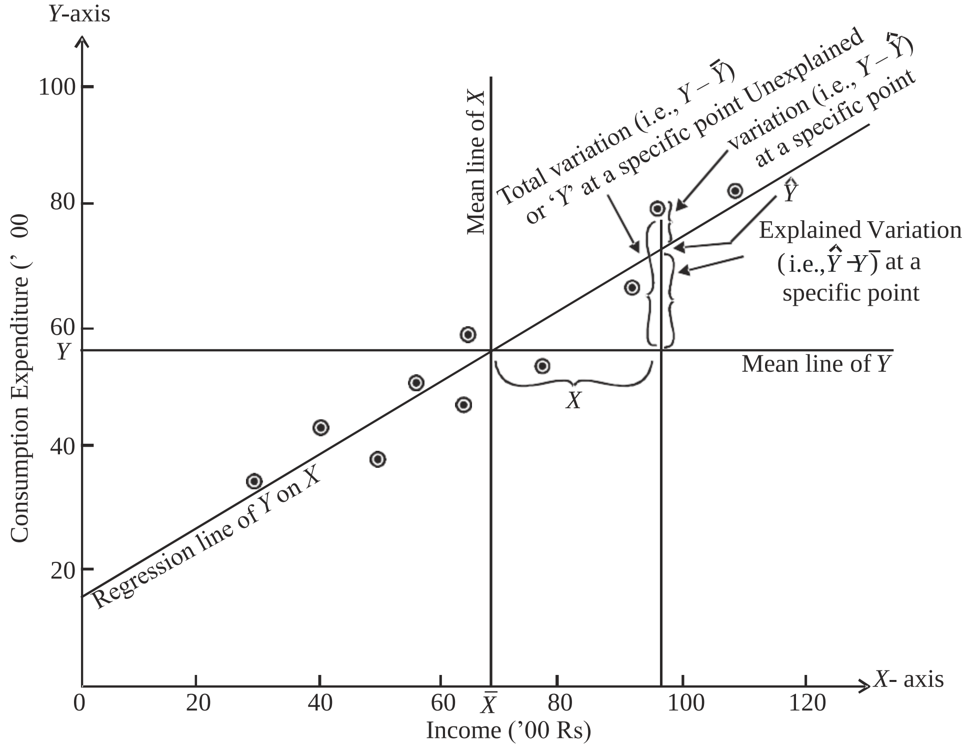


Fig. 9.2 *Diagram showing Total, Explained and Unexplained Variations*

Coefficient of determination is that fraction of the total variation of Y which is explained by the regression line. In other words, coefficient of determination is the ratio of explained variation to total variation in the Y variable related to the X variable. Coefficient of determination algebraically can be stated as under:

$$r^2=\frac{\text{Explained variation}}{\text{Total variation}}$$

$$=$$

Alternatively r^2 can also be stated as under:

$$r^2 = 1 - \frac{\text{Explained variation}}{\text{Total variation}} \qquad = 1 - \frac{\hat{Y} \quad \bar{Y}^{\,2}}{Y \quad \bar{Y}^{\,2}}$$

NOTES

Interpreting r^2

The coefficient of determination can have a value ranging from zero to one. A value of one can occur only if the unexplained variation is zero which simply means that all the data points in the scatter diagram fall exactly on the regression line. For a zero value to occur, $\Sigma(Y-\bar{Y})^2 = \Sigma(Y-\hat{Y})^2$ which simply means that X tells us nothing about Y and hence there is no regression relationship between X and Y variables. Values between 0 and 1 indicate the 'goodness of fit' of the regression line to the sample data. The higher the value of r^2, the better the fit. In other words, the value of r^2 will lie somewhere between 0 and 1. If r^2 has a zero value then it indicates no correlation but if it has a value equal to 1 then it indicates that there is perfect correlation and as such the regression line is a perfect estimator. But in most of the cases the value of r^2 will lie somewhere between these two extremes of 1 and 0. One should remember that r^2 close to 1 indicates a strong correlation between X and Y while an r^2 near zero means there is little correlation between these two variables.

r^2 value can as well be interpreted by looking at the amount of the variation in Y, the dependant variable, that is explained by the regression line. Supposing we get a value of $r^2 = 0.925$ then this would mean that the variations in independent variable (say X) would explain 92.5% of the variation in the dependent variable (say Y). If r^2 is close to 1 then it indicates that the regression equation explains most of the variations in the dependent variable.

Example 9.10: Calculate the coefficient of determination (r^2) using data given in example 1. Analyse the result.

Solution: r^2 can be worked out as shown below:

Since,

$$r^2 = 1 \quad \frac{\text{Unexplained variation}}{\text{Total variation}}$$

$$= 1 \quad \frac{Y \quad \hat{Y}^{\,2}}{Y \quad \bar{Y}^{\,2}}$$

As, $\quad Y \quad \bar{Y}^{\,2} \qquad Y^2 \qquad Y^2 \quad n\bar{Y}^2$, we can write,

$$r^2 = 1 \quad \frac{Y \quad \hat{Y}^{\,2}}{Y^2 \quad n\bar{Y}^2}$$

Calculating and putting the various values, we have the following equation:

$$r^2 = 1 \quad \frac{260.54}{34223 \quad 10 \;\; 56.3^{\,2}} \quad 1 \quad \frac{260.54}{2526.10} \quad 0.897$$

Analysis of Result: The regression equation used to calculate the value of coefficient of determination (r^2) from the sample data shows that about 90% of the variations in consumption expenditure can be explained. In other words, it means that the variations in income explain about 90% of variations in consumption expenditure.

NOTES

4.8 REGRESSION ANALYSIS

The term 'regression' was first used in 1877 by Sir Francis Galton who made a study that showed that the height of children born to tall parents will tend to move back or 'regress' toward the mean height of the population. He designated the word regression as the name of the process of predicting one variable from the other variable. He coined the term multiple regression to describe the process by which several variables are used to predict another. Thus, when there is a well-established relationship between variables, it is possible to make use of this relationship in making estimates and to forecast the value of one variable (the unknown or the dependent variable) on the basis of the other variable/s (the known or the independent variable/s). A banker, for example, could predict deposits on the basis of per capita income in the trading area of bank. A marketing manager may plan his advertising expenditures on the basis of the expected effect on total sales revenue of a change in the level of advertising expenditure. Similarly, a hospital superintendent could project his need for beds on the basis of total population. Such predictions may be made by using regression analysis. An investigator may employ regression analysis to test his theory having the cause and effect relationship. All this explains that regression analysis is an extremely useful tool specially in problems of business and industry involving predictions.

4.8.1 Assumptions in Regression Analysis

While making use of the regression technique for making predictions it is always assumed that:

- There is an actual relationship between the dependent and independent variables.
- The values of the dependent variable are random but the values of the independent variable are fixed quantities without error and are chosen by the experimentor.
- There is a clear indication of direction of the relationship. This means that dependent variable is a function of independent variable. When, for example, we say that advertising has an effect on sales, then we are saying that sales has an effect on advertising.
- The conditions (that existed when the relationship between the dependent and independent variable was estimated by the regression) are the same when the regression model is being used. In other words, it simply means that the relationship has not changed since the regression equation was computed.
- The analysis is to be used to predict values within the range (and not for values outside the range) for which it is valid.

4.8.2 Simple Linear Regression Model

In case of simple linear regression analysis, a single variable is used to predict another variable on the assumption of linear relationship (i.e., relationship of the type defined by $Y = a + bX$) between the given variables. The variable to be predicted is called the dependent variable and the variable on which the prediction is based is called the independent variable.

NOTES

Simple linear regression model[2] (or the Regression Line) is stated as,

Where $Y_i = a + bX_i + e_i$

Y_i is the dependent variable

X_i is the independent variable

e_i is unpredictable random element (usually called as residual or error term)

(a) *a* represent the *Y*-intercept, i.e., the intercept specifies the value of the dependent variable when the independent variable has a value of zero. But this term has practical meaning only if a zero value for the independent variable is possible.

(b) *b* is a constant indicating the slope of the regression line. Slope of the line indicates the amount of change in the value of the dependent variable for a unit change in the independent variable.

If the two constants (viz., *a* and *b*) are known, the accuracy of our prediction of *Y* (denoted by $\hat{Y}$ and read as *Y*-hat) depends on the magnitude of the values of e_i. If in the model, all the e_i tend to have very very large values, then the estimates will not be very good but if these values are relatively small, then the predicted values ( ) will tend to be close to the true values (Y_i).

Estimating the intercept and slope of the regression model (or estimating the regression equation)

The two constants or the parameters, viz., '*a*' and '*b*' in the regression model for the entire population or universe are generally unknown and as such are estimated from sample information. The following are the two methods used for estimation:

$\hat{Y} \quad a \quad bX_i$

(a) Scatter diagram method.

(b) Least squares method.

4.8.3 Scatter Diagram Method

This method makes use of the Scatter diagram also known as Dot diagram.

Scatter diagram is a diagram representing two series with the known variable, i.e., independent variable plotted on the *X*-axis and the variable to be estimated, i.e., dependent variable to be plotted on the *Y*-axis on a graph paper (refer Figure 9.3) to get the following information:

2. Usually the estimate of *Y* denoted by is written as,

on the assumption that the random disturbance to the system averages out or has an expected value of zero (i.e., e = 0) for any single observation. *This regression model is known as the Regression line of Y on X from which the value of Y can be estimated for the given value of X.*

NOTES

Income X (Hundreds of Rupees)	Consumption Expenditure Y (Hundreds of Rupees)
41	44
65	60
50	39
57	51
96	80
94	68
110	84
30	34
79	55
65	48

The scatter diagram by itself is not sufficient for predicting values of the dependent variable. Some formal expression of the relationship between the two variables is necessary for predictive purposes. For the purpose, one may simply take a ruler and draw a straight line through the points in the scatter diagram and this way can determine the intercept and the slope of the said line and then the line can be defined as $\hat{Y} = a + bX_i$ with the help of which we can predict Y for a given value of X. But there are shortcomings in this approach. If, for example, five different persons draw such a straight line in the same scatter diagram, it is possible that there may be five different estimates of a and b, specially when the dots are more dispersed in the diagram. Hence, the estimates cannot be worked out only through this approach. A more systematic and statistical method is required to estimate the constants of the predictive equation. The least squares method is used to draw the best fit line.

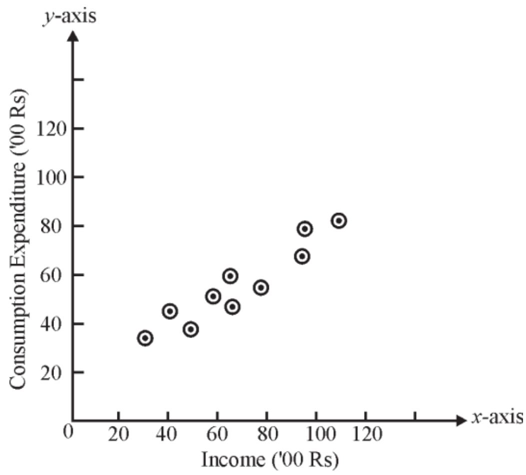


Fig. 9.3 *Scatter Diagram*

4.8.4 Least Squares Method

Least squares method of fitting a line (the line of best fit or the regression line) through the scatter diagram is a method which minimizes the sum of the squared vertical deviations from the fitted line. In other words, the line to be fitted will pass through the points of the scatter diagram in such a way that the sum of the squares of the vertical deviations of these points from the line will be a minimum.

NOTES

The meaning of the least squares criterion can be easily understood through reference to use Figure 9.4 drawn below, where the earlier figure in scatter diagram has been reproduced along with a line which represents the least squares line fit to the data.

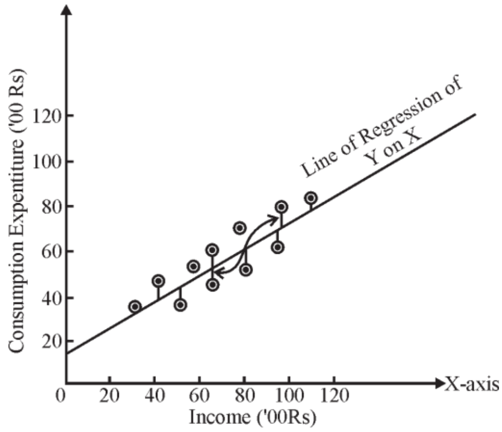


Fig. 9.4 *Scatter Diagram, Regression Line and Short Vertical Lines representing 'e'*

In the above figure the vertical deviations of the individual points from the line are shown as the short vertical lines joining the points to the least squares line. These deviations will be denoted by the symbol '*e*'. The value of '*e*' varies from one point to another. In some cases it is positive, while in others it is negative. If the line drawn happens to be a least squares line, then the values of $\sum e_i$ is the least possible. It is because of this feature the method is known as Least Squares Method.

Why we insist on minimizing the sum of squared deviations is a question that needs explanation. If we denote the deviations from the actual value Y to the estimated value $\hat{Y}$ as $(Y - \hat{Y})$ or e_i, it is logical that we want the $\Sigma(Y - \hat{Y})$ or $\sum_{i=1}^{n} e_i$ to be as small as possible. However, mere examining $\Sigma(Y - \hat{Y})$ or $\sum_{i=1}^{n} e_i$ is inappropriate, since, any e_i can be positive or negative. Large positive values and large negative values could cancel one another. But large values of e_i regardless of their sign, indicate a poor prediction.

Even if we ignore the signs while working out $\sum |e_i|$ the difficulties may continue.

Hence, the standard procedure is to eliminate the effect of signs by squaring each observation. Squaring each term accomplishes two purposes, viz. (*i*) it magnifies (or penalizes) the larger errors, and (*ii*) it cancels the effect of the positive and negative values (since a negative error when squared becomes positive). The choice of minimizing the squared sum of errors rather than the sum of the absolute values implies that there are many small errors rather than a few large errors. Hence, in obtaining the regression line we follow the approach that the sum of the squared deviations be minimum and on this basis work out the values of its constants, viz. '*a*' and '*b*' also known as the intercept and the slope of the line. This is done with the help of the following two normal equations:

$$\Sigma Y = na + b\Sigma X$$
$$\Sigma XY = a\Sigma X + b\Sigma X^2$$

In the above two equations, '*a*' and '*b*' are unknown and all other values, viz. $\sum X$, $\sum Y$, $\sum X^2$, $\sum XY$ are the sum of the products and cross-products to be calculated from the sample data, and '*n*' means the number of observations in the sample.

NOTES

The following examples explain the least squares method.

Example 9.11: Fit a regression line by the method of least squares to the given sample information.

Observations	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
Income (X) ('00 Rs)	41	65	50	57	96	94	110	30	79	65
Consumption Expenditure (Y) ('00 Rs)	44	60	39	51	80	68	84	34	55	48

Solution: We are to fit a regression line $\hat{Y} = a + bX_i$ to the given data by the method of least squares. Accordingly, work out the '*a*' and '*b*' values with the help of the normal equations as stated above and also for the purpose work out $\sum X$, $\sum Y$, $\sum XY$, $\sum X^2$ values from the given sample information table on Summations for Regression Equation.

Summations for Regression Equation

Observations	*Income X ('00 Rs)*	*Consumption Expenditure Y ('00 Rs)*	*XY*	X^2	Y^2
1	41	44	1804	1681	1936
2	65	60	3900	4225	1600
3	50	39	1950	2500	1521
4	57	51	2907	3249	2601
5	96	80	7680	9216	6400
6	94	68	6392	8836	4624
7	110	84	9240	12100	7056
8	30	34	1020	900	1156
9	79	55	4345	6241	3025
10	65	48	3120	4225	2304
$n = 10$	$\sum X = 687$	$\sum Y = 563$	$\sum XY = 42358$	$\sum X^2 = 53173$	$\sum Y^2 = 34223$

Putting the values in the required normal equations we have,

$$563 = 10a + 687b$$

$$42358 = 687a + 53173b$$

Solving these two equations for *a* and *b* we obtain,

$$a = 14.000 \quad \text{and} \quad b = 0.616$$

Hence, the equation for the required regression line is,

$$\hat{Y} = a + bX_i$$

or

$$= 14.000 + 0.616X_i$$

This equation is known as the regression equation of *Y* on *X* from which *Y* values can be estimated for given values of *X* variable.

4.8.5 Checking the Accuracy of Estimating Equation

After finding the regression line as stated above, one can check its accuracy also. The method to be used for the purpose follows from the mathematical property of a line fitted by the method of least squares, viz. the individual positive and negative errors must sum to zero. In other words, using the estimating equation one must find out whether the term $\Sigma(Y - \hat{Y})$ is zero and if this is so, then one can reasonably be sure that he has not committed any mistake in determining the estimating equation.

The Problem of Prediction

When we talk about prediction or estimation, we usually imply that if the relationship $Y_i = a + bX_i + e_i$ exists, then the regression equation $\hat{Y} = a + bX_i$ provides a basis for making estimates of the value for Y which will be associated with particular values of X. In Example 9.11, we worked out the regression equation for the income and consumption data as:

$$\hat{Y} = 14.000 + 0.616X_i$$

On the basis of this equation, we can make a *point estimate* of Y for any given value of X. Suppose, we wish to estimate the consumption expenditure of individuals with income of Rs 10,000. We substitute $X = 100$ for the same in our equation and get an estimate of consumption expenditure as follows:

$$\hat{Y} = 14.000 + 0.616(100) = 75.60$$

Thus, the regression relationship indicates that individuals with Rs 10,000 of income may be expected to spend approximately Rs 7560 on consumption. But this is only an expected or an estimated value and it is possible that actual consumption expenditure of same individual with that income may deviate from this amount and if so, then our estimate will be an error, the likelihood of which will be high if the estimate is applied to any one individual. The *interval estimate* method is considered better and it states an interval in which the expected consumption expenditure may fall. Remember that the wider the interval, the greater the level of confidence we can have, but the width of the interval (or what is technically known as the precision of the estimate) is associated with a specified level of confidence and is dependent on the variability (consumption expenditure in our case) found in the sample. This variability is measured by the standard deviation of the error term 'e', and is popularly known as the standard error of the estimate.

4.8.6 Standard Error of the Estimate

Standard Error (SE) of estimate is a measure developed by the statisticians for measuring the reliability of the estimating equation. Like the standard deviation, the Standard Error (SE) of $\hat{Y}$ measures the variability or scatter of the observed values of Y around the regression line. Standard Error of Estimate (SE of $\hat{Y}$) is worked out as under:

$$\text{SE of } \hat{Y} \text{ (or } S_e) = \sqrt{\frac{\Sigma(Y - \hat{Y})^2}{n - 2}} = \sqrt{\frac{\Sigma e^2}{n - 2}}$$

Where, SE of $\hat{Y}$ (or S_e) = Standard error of the estimate.

Y = Observed value of Y.

$\hat{Y}$ = Estimated value of Y.

e = The error term = $(Y - \hat{Y})$.

n = Number of observations in the sample.

NOTES

Note: In the above formula, $n - 2$ is used instead of n because of the fact that two degrees of freedom are lost in basing the estimate on the variability of the sample observations about the line with two constants, viz., 'a' and 'b' whose position is determined by those same sample observations.

The square of the S_e also known as the variance of the error term is the basic measure of reliability. The larger the variance, the more significant are the magnitudes of the e's and the less reliable is the regression analysis in predicting the data.

4.8.7 Interpreting the Standard Error of Estimate and Finding the Confidence Limits for the Estimate in Large and Small Samples

The larger the SE of estimate (SE_e), the greater happens to be the dispersion or scattering of the given observations around the regression line. But if the SE of estimate happens to be zero, then the estimating equation is a 'perfect' estimator (i.e., 100 per cent correct estimator) of the dependent variable.

In case of large samples, i.e., where $n > 30$ in a sample, it is assumed that the observed points are normally distributed around the regression line and we may find,

68% of all points within SE_e limits

95.5% of all points within SE_e limits

99.7% of all points within SE_e limits

This can be stated as:

4.8.7.1 The observed values of Y are normally distributed around each estimated value of

, and

4.8.7.2 The variance of the distributions around each possible value of is the same.

In case of small samples, i.e., where $n \le 30$ in a sample, the 't' distribution is used for finding the two limits more appropriately.

This is done as follows:

Upper limit = + 't' (SE_e)

Lower limit = – 't' (SE_e)

Where, = The estimated value of Y for a given value of X.

SE_e = The standard error of estimate.

't' = Table value of 't' for given degrees of freedom for a specified confidence level.

4.8.8 Some Other Details concerning Simple Regression

Sometimes the estimating equation of Y, also known as the Regression equation of Y on X, is written as follows:

=

Or, $\hat{Y}$ =

Where, r = Coefficient of simple correlation between X and Y

$\sigma_Y =$ Standard deviation of Y

NOTES

σ_X = Standard deviation of X

$\bar{X}$ = Mean of X

$\bar{Y}$ = Mean of Y

$\hat{Y}$ = Value of Y to be estimated

X_i = Any given value of X for which Y is to be estimated.

This is based on the formula we have used, i.e., . The coefficient of X_i is defined as,

$$\text{Coefficient of } X_i = b \ = \ r\frac{Y}{X}$$

Also known as regression coefficient of Y on X or slope of the regression line of Y on X or b_{YX}.

$$= \qquad\qquad = \frac{XY \quad n\bar{X}\bar{Y}}{X^2 \quad n\bar{X}^2}$$

And, $\qquad a = r\frac{Y}{X}\ \bar{X} \quad \bar{Y}$ $\qquad \left(\text{Since } b = r\frac{\sigma_Y}{\sigma_X}\right)$

$$= \bar{Y} \quad b\bar{X}$$

Similarly, the estimating equation of X also known as the regression equation of X on Y can be stated as:

$$\hat{X} \quad \bar{X} \quad = r\frac{X}{Y}\ Y \quad \bar{Y}$$

$\hat{X}\ b_{XY}\ r\frac{\sigma_X}{\sigma_Y}\ n\bar{X}\bar{Y} \quad \sqrt{Y^2 \quad n\bar{Y}^2}$ Or, $\sqrt{Y^2 \quad n\bar{Y}^2}\sqrt{X^2 \quad n\bar{X}^2}\sqrt{X^2 \quad n\bar{X}^2}$ $\qquad \hat{X} =$

And the Regression coefficient of X on Y (or b_{XY}) $\quad r\frac{X}{Y} \quad \frac{XY \quad n\bar{X}\bar{Y}}{Y^2 \quad n\bar{Y}^2}$

If we are given the two regression equations as stated above along with the values of 'a' and 'b' constants to solve the same for finding the value of X and Y, then the values o f X and Y so obtained are the mean value of X (i.e., $\bar{X}$) and the mean value of Y (i.e., $\bar{Y}$).

If we are given the two regression coefficients (viz., b_{XY} and b_{YX}) then we can work out the value of coefficient of correlation by just taking the square root of the product of the regression coefficients as shown below:

$$r =$$

$$= \qquad\qquad = \qquad = r$$

The (±) sign of r will be determined on the basis of the sign of the regression coefficients given. If regression coefficients have minus signs then r will be taken with minus (–) sign and if regression coefficients have plus signs then r will be taken with plus (+) sign. Remember that both regression coefficients will necessarily have the same sign whether it is minus or plus for their sign is governed by the sign of coefficient of correlation.

Example 9.12: Given is the following information:

Mean	39.5	47.5
Standard Deviation	10.8	17.8

Simple correlation coefficient between X and Y is = + 0.42

Find the estimating equation of Y and X.

Solution: Estimating equation of Y can be worked out as,

□ =

Or $\hat{Y}$ =

=

$= 0.69X_i \quad 27.25 \quad 47.5$

$= 0.69X_i + 20.25$

Similarly, the estimating equation of X can be worked out as under:

$$\hat{X} \quad \bar{X} = r\frac{\sigma_X}{\sigma_Y}\left(Y_i - \bar{Y}\right)$$

Or, =

Or, =

$= 0.26Y_i - 12.25 + 39.5$

$= 0.26Y_i + 27.25$

Example 9.13: Given is the following data:

Variance of $X = 9$

Regression equations:

$$4X - 5Y + 33 = 0$$
$$20X - 9Y - 107 = 0$$

Find (a) Mean values of X and Y

(b) Coefficient of Correlation between X and Y

(c) Standard deviation of Y

Solution: (a) For finding the mean values of X and Y we solve the two given regression equations for the values of X and Y as follows:

$4X - 5Y + 33 = 0$ (1)

$20X - 9Y - 107 = 0$ (2)

If we multiply Equation (1) by 5 we have the following equations:

$$20X - 25Y = -165 \qquad (3)$$

$$20X - 9Y = 107 \qquad (2)$$

$$- \quad + \quad -$$

$$-16Y = -272 \quad \text{Subtracting Equation (2) from (3)}$$

Or, $Y = 17$

Putting this value of Y in Equation (1) we have,

$$4X = -33 + 5(17)$$

Or, $X = \frac{33 \quad 85}{4} \quad \frac{52}{4} \quad 13$

Hence, $\bar{X} = 13$ and $= 17$

(b) For finding the coefficient of correlation, first of all we presume one of the two given regression equations as the estimating equation of X. Let equation $4X - 5Y + 33 = 0$ be the estimating equation of X, then we have,

From this we can write $b_{XY} \quad \frac{5}{4}$

The other given equation is then taken as the estimating equation of Y and can be written as,

4 20X 33 And from this we can write $b_{YX} \quad \frac{20}{9}$. $\hat{X}$ $\hat{Y}$ 5 Y 0.2 X 0.6 49 3 X 4 9

If the above equations are correct then r must be equal to

$$r = \sqrt{5/4 \quad 20/9} \quad \sqrt{25/9} = 5/3 = 1.6$$

which is an impossible equation, since r can in no case be greater than 1. Hence, we change our supposition about the estimating equations and by reversing it, we re-write the estimating equations as under:

$$\hat{X} \quad \frac{9Y_i}{20} \quad \frac{107}{20} \quad \text{and} \quad \hat{Y} \quad \frac{4X_i}{5} \quad \frac{33}{5}$$

Hence, $r = \sqrt{9/20 \quad 4/5} = \sqrt{9/25}$

$= 3/5 \qquad = 0.6$

Since, regression coefficients have plus signs, we take $r = +0.6$

(c) Standard deviation of Y can be calculated as follows:

 Variance of $X = 9$ $\therefore$ Standard deviation of $X = 3$

 =

Hence, $\sigma_Y = 4$

Alternatively, we can work it out as under:

□ =

Hence, $\sigma_Y = 4$

NOTES

4.9 SUMMARY

- Correlation analysis is the statistical tool generally used to describe the degree to which one variable is related to another. The relationship, if any, is usually assumed to be a linear one. This analysis is used quite frequently in conjunction with regression analysis to measure how well the regression line explains the variations of the dependent variable.
- Typically, the word correlation refers to the relationship or interdependence between two variables. There are various phenomena which have relation to each other. The theory by means of which quantitative connections between two sets of phenomena are determined is called the 'Theory of Correlation'.
- On the basis of the theory of correlation one can study the comparative changes occurring in two related phenomena and their cause-effect relation can be examined. Thus, correlation is concerned with the relationship between two related and quantifiable variables.
- For correlation, it is essential that the two phenomena should have a cause-effect relationship. If such relationship does not exist then there can be no correlation.
- Correlation can either be linear or it can be non-linear. The non-linear correlation is also known as curvilinear correlation. The distinction is based upon the constancy of the ratio of change between the variables.
- The study of correlation for two variables (of which one is independent and the other is dependent) involves application of simple correlation.
- Statisticians have developed two measures for describing the correlation between two variables, viz., the coefficient of determination and the coefficient of correlation.
- The scatter diagram is a graph of observed plotted points where each point represents the values of *X* and *Y* as a coordinate. It portrays the relationship between these two variables graphically.
- The line of best fit is known as the regression line and the algebraic expression that identifies this line is a general straight line equation and is given as, $Y_c = b_0 + b_1X$, where b_0 and b_1 are the two pieces of information called parameters which determine the position of the line completely. Parameter b_0 is known as the *Y*-intercept (or the value of Y_c at $X = 0$) and parameter b_1 determines the slope of the regression line which is the change in Y_c for each unit change in *X*.
- The measure, standard error of the estimate is used to determine the dispersion of observed values of *Y* about the regression line.
- The coefficient of correlation symbolically denoted by '*r*' is another important measure to describe how well one variable is explained by another. It measures the degree of relationship between the two causally-related variables. The value of this coefficient can never be more than +1 or less than –1. Thus +1 and –1 are the limits of this coefficient.

Check Your Progress

5. What is coefficient of determination (r^2)?
6. What is regression analysis?
7. What are the types of constants involved in regression?
8. What are the types of methods to calculate the constants in regression models?
9. Can the two regression lines coincide?

- If the coefficient of correlation has a zero value then it means that there exists no correlation between the variables under study.
- Karl Pearson's method is the most widely-used method of measuring the relationship between two variables. There is a linear relationship between the two variables which means that straight line would be obtained if the observed data are plotted on a graph.
- Probable Error (PE) of r is very useful in interpreting the value of r and is worked out as under for Karl Pearson's coefficient of correlation:

$$\text{PE} = 0.6745\ \frac{1-r^2}{\sqrt{n}}$$

- If r is less than its PE, it is not at all significant. If r is more than PE, there is correlation. If r is more than 6 times its PE and greater than ± 0.5, then it is considered significant.
- The coefficient of determination (symbolically indicated as r^2, though some people would prefer to put it as R^2) is a measure of the degree of linear association or correlation between two variables, say X and Y, one of which happens to be an independent variable and the other a dependent variable.
- If we subtract the unexplained variation from the total variation, we obtain the explained variation, i.e., the variation explained by the line of regression. Thus, Explained Variation = (Total variation) – (Unexplained variation).
- Coefficient of determination is that fraction of the total variation of Y which is explained by the regression line. Coefficient of determination algebraically can be stated as under:

$$r^2 = \frac{\text{Explained variation}}{\text{Total variation}}$$

- The coefficient of determination can have a value ranging from zero to one. A value of one can occur only if the unexplained variation is zero which simply means that all the data points in the scatter diagram fall exactly on the regression line.
- Values between 0 and 1 indicate the 'goodness of fit' of the regression line to the sample data. The higher the value of r^2, the better the fit.
- The term 'regression' was first used in 1877 by Sir Francis Galton who made a study that showed that the height of children born to tall parents will tend to move back or 'regress' toward the mean height of the population.
- Sir Francis Galton designated the word regression as the name of the process of predicting one variable from the other variable. He coined the term multiple regression to describe the process by which several variables are used to predict another.
- In case of simple linear regression analysis, a single variable is used to predict another variable on the assumption of linear relationship (i.e., relationship of the type defined by $Y = a + bX$) between the given variables. The variable to be predicted is called the dependent variable and the variable on which the prediction is based is called the independent variable.

NOTES

- The two constants or the parameters, viz., '*a*' and '*b*' in the regression model for the entire population or universe are generally unknown and as such are estimated from sample information. The two methods used for estimation are (a) Scatter diagram method and (b) Least squares method.
- Scatter diagram method is also known as Dot diagram. It represents two series with the known variable, i.e., independent variable plotted on the *X*-axis and the variable to be estimated, i.e., dependent variable to be plotted on the *Y*-axis.
- Least squares method of fitting a line (the line of best fit or the regression line) through the scatter diagram is a method which minimizes the sum of the squared vertical deviations from the fitted line. In other words, the line to be fitted will pass through the points of the scatter diagram in such a way that the sum of the squares of the vertical deviations of these points from the line will be a minimum.
- Standard Error (SE_e) of estimate is a measure developed by the statisticians for measuring the reliability of the estimating equation. The larger the SE of estimate (SE_e), the greater be the dispersion or scattering of the given observations around the regression line.
- The (±) sign of *r* will be determined on the basis of the sign of the regression coefficients given. If regression coefficients have minus signs then *r* will be taken with minus (–) sign and if regression coefficients have plus signs then *r* will be taken with plus (+) sign.

4.10 KEY TERMS

- **Correlation:** Relationship or interdependence between two variables
- **Correlation analysis:** The statistical tool used to describe the degree to which one variable is related to another
- **Multiple correlation:** The relationship between a dependent variable and two or more independent variables
- **Scatter diagram:** A diagram representing two series with the known variables, i.e., independent variable plotted on the *X*-axis and the variable to be estimated, i.e., dependent variable to be plotted on the *Y*-axis on a graph for the given information
- **Rank correlation:** A descriptive index of agreement between the ranks over individuals
- **Regression analysis:** The relationship used for making estimates and forecasts about the value of one variable (the unknown or the dependent variable) on the basis of the other variable/s (the known or the independent variable/s)
- **Standard error of the estimate:** The measure developed by statisticians for measuring the reliability of the estimating equation

4.11 ANSWERS TO 'CHECK YOUR PROGRESS'

1. The types of correlations are:
   (a) Positive or negative correlations
   (b) Linear or non-linear correlations
   (c) Simple, partial or multiple correlations

2. Scatter diagram is a method to calculate the constants in regression models that makes use of scatter diagram or dot diagram. A scatter diagram is a diagram that represents two series with the known variables, i.e., independent variable plotted on the *X*-axis and the variable to be estimated, i.e., dependent variable to be plotted on the *Y*-axis.
3. The coefficient of correlation, which is symbolically denoted by *r*, is an important measure to describe how well one variable explains another. It measures the degree of relationship between two causally-related variables. The value of this coefficient can never be more than +1 or –1. Thus, +1 and –1 are the limits of this coefficient.
4. The rank correlation, written *r*, is a descriptive index of agreement between ranks over individuals. It is the same as the ordinary coefficient of correlation computed on ranks, but its formula is simpler.
5. The coefficient of determination (r^2), the square of the coefficient of correlation (*r*), is a precise measure of the strength of the relationship between the two variables and lends itself to more precise interpretation because it can be presented as a proportion or as a percentage.
6. Regression analysis is an extremely useful tool especially in problems of business and industry for making predictions. A banker, for example, could predict deposits on the basis of per capita income in the trading area of bank. A marketing manager may plan his advertising expenditures on the basis of the expected effect on total stales revenue of a change in the level of advertising expenditure, and so on.
7. The two constants involved in regression model are *a* and *b*, where *a* represents the *Y*-intercept and *b* indicates the slope of the regression line.
8. There are two methods to calculate the constants in regression models. They are:
   (a) Scatter diagram method (b) Least squares method
9. Two regression lines can coincide if and only if all the points in the scatter diagram lie on one straight line, i.e., if the correlation is perfect, $r = 1$.

4.12 QUESTIONS AND EXERCISES

Short-Answer Questions

1. How you will predict the value of dependent variable?
2. Differentiate between scatter diagram and least squares method.
3. Can the accuracy of estimated equation be checked? How?
4. How the standard error of estimate is calculated?
5. What is the importance of correlation analysis?
6. How will you determine the coefficient of determination?
7. How is the least squares method useful in statistical calculations?
8. How does the scatter diagram help in studying correlation between two variables?
9. Write the method for calculating the coefficient of correlation by Karl Pearson's method.
10. Define the term regression analysis.
11. What is concurrent deviation?

12. Write the formulae for calculating coefficient of concurrent deviation.
13. If deviation of two time series is concurrent, what will be the nature of graph?

Long-Answer Questions

1. Explain the meaning and significance of regression and correlation analysis.
2. What is a 'Scatter diagram'? How does it help in studying correlation between two variables? Explain with the help of examples.
3. Obtain the estimating equation by the method of least squares from the following information:

X (*Independent Variable*)	*Y* (*Dependent Variable*)
2	18
4	12
5	10
6	8
8	7
11	5

4. Calculate correlation coefficient from the following results:

$$n = 10;\ \Sigma X = 140;\ \Sigma Y = 150$$
$$\Sigma(X - 10)^2 = 180;\ \Sigma(Y - 15)^2 = 215$$
$$\Sigma(X - 10)(Y - 15) = 60$$

5. Given is the following information:

Observation	*Test Score* *X*	*Sales* ('000 Rs) *Y*
1	73	450
2	78	490
3	92	570
4	61	380
5	87	540
6	81	500
7	77	480
8	70	430
9	65	410
10	82	490
Total	766	4740

On the basis of above information,

(i) Graph the scatter diagram for the above data.

(ii) Find the regression equation $\hat{Y} = a + bX_i$ and draw the line corresponding to the equation on the scatter diagram.

(iii) On the basis of calculated values of the coefficients of regression equation analyse the relationship between test scores and sales.

(iv) Make an estimate about sales if the test score happens to be 75.

6. As a furniture retailer in a certain locality, you are interested in finding the relationship that might exist between the number of building permits issued in that locality in past years and the volume of your sales in those years. You accordingly collected

the data for your sales (Y, in thousands of rupees) and the number of building permits issued (X, in hundreds) in the past 10 years. The results worked out as:

$$n = 10, \sum X = 200, \sum Y = 2200$$

$$\sum X^2 = 4600, \sum XY = 45800, \sum Y^2 = 490400$$

Answer the following:

(i) Calculate the coefficients of the regression equation.

(ii) It is expected that there will be approximately 2000 building permits to be issued next year. On this basis, what level of sales can you expect next year?

(iii) On the basis of the relationship you found in (a) one would expect what change in sales with an increase of 100 building permits?

(iv) State your estimate of (b) in the (c) so that the level of confidence you place in it is 0.90.

7. Are the following two statements consistent? Give reasons for your answer.

(a) The regression coefficient of X on Y is 3.2

(b) The regression coefficient of Y on X is 0.8

8. Regression of savings (S) of a family on income (Y) may be expressed as $S \quad a \quad \frac{Y}{m}$, where '$a$' and '$m$' are constants. In random sample of 100 families, the variance of savings is one-quarter of the variance of incomes and the coefficient of correlation is found to be +0.4. Obtain the estimate of 'm'.

9. Calculate correlation coefficient and the two regression lines for the following information:

		Ages of Wives (in years)				
		10–20	*20–30*	*30–40*	*40–50*	*Total*
Ages of	10–20	20	26	—	—	46
Husbands	20–30	8	14	37	—	59
(in	30–40	—	4	18	3	25
years)	40–50	—	—	4	6	10
	Total	28	44	59	9	140

10. Two random variables have the regression with equations,

$$3X + 2Y - 26 = 0$$

$$6X + Y - 31 = 0$$

Find the mean value of X as well as of Y and the correlation coefficient between X and Y. If the variance of X is 25, find σ_Y from the data given above.

11. (*i*) Give one example of a pair of variables which would have,

- An increasing relationship
- No relationship
- A decreasing relationship

- Suppose that the general relationship between height in inches (X) and weight in kg (Y) is $Y' = 10 + 2.2\,(X)$. Consider that weights of persons of a given height are normally distributed with a dispersion measurable by $\sigma_e = 10$ kg.

NOTES

- What would be the expected weight for a person whose height is 65 inches?
- If a person whose height is 65 inches should weigh 161 kg., what valueof *e* does this represent?
- What reasons might account for the value of *e* for the person in case (*ii*)?
- What would be the probability that someone whose height is 70 inches would weigh between 124 and 184 kg?

12. Calculate correlation coefficient from the following results:

$$n = 10;\ \Sigma X = 140;\ \Sigma Y = 150$$
$$\Sigma(X - 10)^2 = 180;\ (Y - 15)^2 = 215$$
$$\Sigma(X - 10)\ (Y - 15)) = 60$$

13. Examine the following statements and state whether each one of the statements is true or false, assigning reasons to your answer.
    (i) If the value of the coefficient of correlation is 0.9 then this indicates that 90% of the variation in dependent variable has been explained by variation in the independent variable.
    (ii) It would not be possible for a regression relationship to be significant if the value of r^2 was less than 0.50.
    (iii) If there is found a high significant relationship between the two variables *X* and *Y*, then this constitutes definite proof that there is a casual relationship between these two variables.
    (iv) Negative value of the '*b*' coefficient in a regression relationship indicates a weaker relationship between the variables involved than would a positive value for the '*b*' coefficient in a regression relationship.
    (v) If the value for the '*b*' coefficient in an estimating equation is less than 0.5, then the relationship will not be a significant one.
    (vi) $r^2 + k^2$ is always equal to one. From this it can also be inferred that $r + k$ is equal to one

14. Find coefficient of correlation from the following data by the concurrent deviation method:

15. Following table is given for two series. Find the coefficient of correlation by concurrent deviation method.

X:	80	78	75	75	58	67	60	59
Y:	12	13	14	14	14	16	15	17

16. A student appears for his mathematics and science examinations in 8 unit tests and his percentage score is noted as below. Find the coefficient of correlation by concurrent deviation method.

Maths	90	93	89	86	90	90	91	92
Science	88	89	85	87	87	90	91	88

NOTES

4.13 FURTHER READING

Allen, R.G.D. 2008. *Mathematical Analysis For Economists*. London: Macmillan and Co., Limited.

Chiang, Alpha C. and Kevin Wainwright. 2005. *Fundamental Methods of Mathematical Economics*, 4 edition. New York: McGraw-Hill Higher Education.

Yamane, Taro. 2012. *Mathematics For Economists: An Elementary Survey*. USA: Literary Licensing.

Baumol, William J. 1977. *Economic Theory and Operations Analysis*, 4th revised edition. New Jersey: Prentice Hall.

Hadley, G. 1961. *Linear Algebra*, 1st edition. Boston: Addison Wesley.

Vatsa , B.S. and Suchi Vatsa. 2012. *Theory of Matrices*, 3rd edition. London: New Academic Science Ltd.

Madnani, B C and G M Mehta. 2007. *Mathematics for Economists*. New Delhi: Sultan Chand & Sons.

Henderson, R E and J M Quandt. 1958. *Microeconomic Theory: A Mathematical Approach*. New York: McGraw-Hill.

Nagar, A.L. and R.K.Das. 1997. *Basic Statistics*, 2nd edition. United Kingdom: Oxford University Press.

Gupta, S.C. 2014. *Fundamentals of Mathematical Statistics*. New Delhi: Sultan Chand & Sons.

M. K. Gupta, A. M. Gun and B. Dasgupta. 2008. *Fundamentals of Statistics*. West Bengal: World Press Pvt. Ltd.

Saxena, H.C. and J. N. Kapur. 1960. *Mathematical Statistics*, 1st edition. New Delhi: S. Chand Publishing.

Hogg, Robert V., Joeseph McKean and Allen T Craig. *Introduction to Mathematical Statistics*, 7th edition. New Jersey: Pearson.

NOTES

UNIT V INDEX NUMBER AND TIME SERIES

Structure

5.0 INTRODUCTION

In this unit, you will learn about index numbers, which refer to a specialized type of average. Index numbers have a wide application, including industry, agriculture and business. The unit also discusses methods of constructing index numbers. You will also learn the different types of index numbers, such as weighted index numbers, volume index numbers and value index numbers. Moreover, this unit will also familiarize you with the uses and importance of index numbers.

You will also learn how time series analysis differs from regression analysis. We often see a number of charts on company drawing boards or in newspapers, where we see lines going up and down from left to right on a graph. The vertical axis represents a variable, such as productivity or crime data in the city, and the horizontal axis represents the different periods of increasing time, such as days, weeks, months or years. The analysis of the movements of such variables over periods of time is referred to as time series analysis. Time series can then be defined as a set of numeric observations of the dependent variable, measured at specific points in time in a chronological order, usually at equal intervals, in order to determine the relationship of time to such variables.

5.1 UNIT OBJECTIVES

After going through this unit, you will be able to:

- Understand the meaning of index numbers
- Explain the different types of index numbers
- Describe the uses and importance of index numbers
- Classify the time series

NOTES

- Analyse the components of time series
- Describe the influence of time series analysis
- Explain the different methods of measuring trend
- Understand the different methods of measuring seasonal variations
- Explain the significance smoothing techniques
- Calculate simple averages and moving averages
- Analyse exponential smoothing
- Measure irregular variations and seasonal adjustments

5.2 MEANING AND IMPORTANCE OF INDEX NUMBERS

Index numbers are a specialized type of average. They are designed to measure the relative change in the level of a phenomenon with respect to time, geographical locations or some other characteristics. As we have seen that averages are used to compare two or more series as they represent their central tendencies. But there is a great limitation in the use of averages. They can be used to compare only those series which are expressed in the same units. But if the units in which two or more series are expressed are different, or if the series are composed of different types of items, averages cannot be used to compare them. For instance, if we want to measure the relative change in the price level, we shall not be able to do so by using the averages because prices of different commodities are expressed in different units, such as per metre, per kilogram, per metric ton, etc. In such cases, we require some special type of average which will enable us to measure changes in the price level. Index numbers are such an average. According to Wheldon, '*Index number is a statistical device for indicating the relative movements of the data where measurement of actual movements is difficult or incapable of being made*.' According to F.Y. Edgeworth, '*Index number shows by its variations the changes in a magnitude which is not susceptible either of accurate measurement in itself or of direct valuation in practice*.'

Originally, the index numbers were developed for measuring the effect of changes in the price level. But today the index numbers are also used to measure changes in industrial production, fluctuations in the level of business activities or variations in the agricultural output, etc. In fact, if we want to get an idea as to what is happening to an economy, we have simply to look to a few important indices like those of industrial output, agricultural production and business activity. In the words of G. Simpson and F. Kafka: '*Index numbers are today one of the most widely used statistical devices. They are used to take the pulse of the economy and they have come to be used as indicators of inflationary or deflationary tendencies*.'

5.3 TYPES OF INDEX NUMBERS

Methods of constructing index numbers can broadly be divided into two classes namely:

(a) Unweighted indices

(b) Weighted indices

NOTES

In case of unweighted indices, weights are not expressly assigned, whereas in the weighted indices weights are expressly assigned to the various items. Each of these types may be further classified under two heads:

(i) Aggregate of prices method

(ii) Average of price relatives method

Figure 10.1 illustrates the various methods of constructing index numbers:

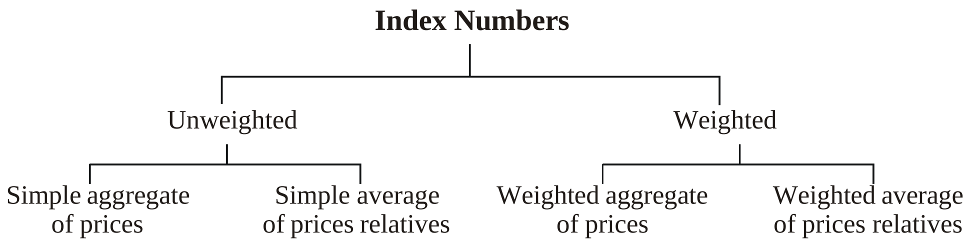


Fig. 10.1 *Methods of Constructing Index Numbers*

A. Unweighted Index Numbers

(1) Simple Aggregate of Prices Method

Under this method, the total of prices for all commodities in the current year is divided by the total of prices for these commodities in the base year and the quotient is multiplied by 100. Symbolically,

$$P_{01} = \frac{\sum P_1}{\sum P_0} \times 100$$

where,

$\sum P_1$ = Total of current year prices for various commodities.

$\sum P_0$ = Total of base year prices for various commodities.

This method of constructing index numbers is very simple and requires the following steps for its computation:

(i) Total the prices of various commodities for each time period to get $\sum P_0$ and $\sum P_1$. These totals are in rupees.

(ii) Divide the total of the given time period, $\sum P_1$, by the base period total, $\sum P_0$ and express the result in per cent, by multiplying the quotient by 100.

Example 10.1: From the following data, construct an index number of prices by simple aggregative method for 1982 taking 1981 as the base:

Commodity	*Unit*	*Price in 1981*	*Price in 1982*
Milk	litre	2.00	2.50
Butter	kg	12.00	15.00
Cheese	kg	10.00	12.00
Bread	One	2.00	2.50
Eggs	dozen	4.00	5.00

NOTES

Solution: Construction of index numbers

Commodity	*Unit*	P_0	P_1
Milk	litre	2.00	2.50
Butter	kg	12.00	15.00
Cheese	kg	10.00	12.00
Bread	One	2.00	2.50
Eggs	dozen	4.00	5.00
		$\Sigma P_0 = 30.00$	$\Sigma P_1 = 37.00$

$$P_{01} = \frac{\sum P_1}{\sum P_0} \times 100 = \frac{37}{30} \times 100 = 123.33\%$$

This means that as compared to 1981, there is a net increase of 123.33 per cent in 1982, in the prices of commodities included in the index.

This method suffers from two drawbacks, which are as follows:

(i) The unit by which each item is priced introduces a concealed weight in the simple aggregate of actual prices. For instance, milk is quoted per litre in example 1. If the price is expressed in terms of per gallon, the index might be very different.

(ii) Equal weightage is given to all the items irrespective of their relative importance.

2. Simple Average of Price Relative Method

Under this method, the price relatives for each commodity are calculated and their average is found out. The steps involved in the construction of this index are as follows:

(i) Obtain the price relative by dividing the price of each commodity in the given time period, P_1 by its price in the base period, P_0 and express this result in per cent, i.e., obtain $\frac{P_1}{P_0} \times 100$ for each commodity.

(ii) Average these price relatives for the given time period by dividing the total of price relatives for different commodities by the number of commodities. Symbolically,

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

where N refers to the number of commodities (items) whose price relatives are thus averaged.

Example 10.2: From the data given in Example 10.1, compute the price index for 1982 with 1981 as base, by simple average of price relatives method.

NOTES

Solution: Construction of price index

Commodities	*Unit*	*Price in 1981* P_0 (`)	*Price in 1982* P_1 (`)	*Price Relative*
Milk	litre	2.00	2.50	
Butter	kg	12.00	15.00	$\frac{15}{12} \times 100 = 125$
Cheese	kg	10.00	12.00	$\frac{12}{10} \times 100 = 120$
Bread	one	2.00	2.50	$\frac{2.50}{2.00} \times 100 = 125$
Eggs	dozen	4.00	5.00	$\frac{5}{4} \times 100 = 125$
$N = 5$				$\Sigma\left(\frac{P_1}{P_0} \times 100\right) = 620$

$$P_{01} = \frac{\Sigma\left(\frac{P_1}{P_0} \times 100\right)}{N} = \frac{620}{5} = 124$$

The simple average of price relatives method is superior to the simple aggregate of prices method in two respects:

$\frac{2.50}{2.00} \times 100 = 125$

(i) Since we are comparing price per litre with price per litre, and price per kilogram with price per kilogram, the concealed weight due to use of different units is completely removed.

(ii) The index is not influenced by extreme items as equal importance is given to all items.

However, the greatest drawback of unweighted indices is that equal importance or weight is given to all items included in the index number which is not proper. As such, unweighted indices are of little use in practice.

B. Weighted Index Numbers

1. Weighted Aggregate of Prices Index

These indices are similar to the simple aggregative type with the fundamental difference that weights are assigned explicitly to the various items included in the index. In the matter of assigning weights, authors differ. As a result, a large number of formulae methods have been devised for constructing index numbers. Some of the important formulae methods are as follows:

(i) Laspeyre's Method: In this method, base year quantities are taken as weights. The formula for constructing the index is:

$$P_{01} = \frac{\Sigma P_1 q_0}{\Sigma P_0 q_0} \times 100$$

where P_1 = Price in the current year.

P_0 = Price in the base year.

q_0 = Quantity in the base year.

NOTES

According to this method, the index number for each year is obtained in following three steps:

(a) The price of each commodity in each year is multiplied by the base year quantity of that commodity. For the base year, each product is symbolized by P_0q_0, and for the current year by P_1q_0.

(b) The products for each year are totalled and ΣP_1q_0 and ΣP_0q_0 are obtained.

(c) ΣP_1q_0 is divided by ΣP_0q_0 and the quotient is multiplied by 100 to obtain the index.

Example 10.3: From the following data, calculate the index number of prices for 1982 with 1972 as base using the Laspeyre's method.

	1972		*1982*	
Item	*Price*	*Quantity*	*Price*	*Quantity*
A	2	8	4	6
B	5	10	6	5
C	4	14	5	10
D	2	19	2	13

Solution: Representing base year (1972) price by P_0, base year quantity by q_0, current year (1982) price by P_1 and current year quantity by q_1 we have:

Commodity	P_0	q_0	P_1	q_1	$P_0\,q_0$	$P_1\,q_0$
A	2	8	4	6	16	32
B	5	10	6	5	50	60
C	4	14	5	10	56	70
D	2	19	2	13	38	38
					$\sum P_0q_0 = 160$	$\sum P_1q_0 = 200$

Index number of prices by Laspeyre's method $= \dfrac{\sum P_1q_0}{\sum P_0q_0}\times 100$

$$= \frac{200}{160}\times 100 = 125$$

Laspeyre's index is very widely used. It tells us about the change in the aggregate value of the base period list of goods when valued at a given period price.

However, this index has one drawback. It does not take into consideration the changes in the consumption pattern that take place with the passage of time.

(ii) Paasche's Index : In this method, the current year quantities (q_1) are taken as weights. The formula for constructing this index is:

$$P_{01} = \frac{\sum P_1q_1}{\sum P_0q_1}\times 100$$

NOTES

Steps for constructing the Paasche's index are the same as those taken in constructing Laspeyre's index with the only difference that the price of each commodity in each year is multiplied by the quantity of that commodity in the current year rather than by the quantity in the base year.

Example 10.4: Taking the data given in Example 10.3, compute the index number of prices for 1982 with 1972 as base, using the Paasche's method.

Solution: Construction of Paasche's Index

Commodity	P_0	q_0	P_1	q_1	$P_0\, q_1$	$P_1\, q_1$
A	2	8	4	6	12	24
B	5	10	6	5	25	30
C	4	14	5	10	40	50
D	2	19	2	13	26	26
					$\sum P_0 q_1 =$ 103	$\sum P_1 q_1 =$ 130

Index number of prices by Paasche's method $= \dfrac{\sum P_1 q_1}{\sum P_0 q_1} \times 100$

$$= \frac{130}{103} \times 100 = 126.21$$

Although this method takes into consideration the changes in the consumption pattern, the need for collecting data regarding quantities for each year or each period makes the method very expensive. Hence, where the number of commodities is large, Paasche's method is not preferred.

(iii) Bowley-Drobisch Method: This method is the simple arithmetic mean of Laspeyre's and Paasche's indices. The formula for constructing Bowley-Drobisch index is:

$$P_{01} = \frac{\dfrac{\sum P_1 q_0}{\sum P_0 q_0} + \dfrac{\sum P_1 q_1}{\sum P_0 q_1}}{2} \times 100$$

$$P_{01} = \frac{L + P}{2}$$

Where L = Laspeyre's index

P = Paasche's index

Example 10.5: Compute the index number of prices for 1976 with 1970 as base using the Bowley-Drobisch method from the following data.

	1970		*1976*	
Items	*Price*	*Quantity*	*Price*	*Quantity*
1	2	20	5	15
2	4	4	8	5
3	1	10	2	12
4	5	5	10	6

NOTES

Solution: Computation of price index by Bowley-Drobisch formula,

Items	P_0	q_0	P_1	q_1	P_0q_0	P_0q_1	P_1q_0	P_1q_1
1	2	20	5	15	40	30	100	75
2	4	4	8	5	16	20	32	40
3	1	10	2	12	10	12	20	24
4	5	5	10	6	25	30	50	60
					$\Sigma P_0q_0 = 91$	$\Sigma P_0q_1 = 92$	$\Sigma P_1q_0 = 202$	$\Sigma P_1q_1 = 199$

According to Bowley-Drobisch formula:

$$= \frac{\frac{202}{91} + \frac{199}{92}}{2} \times 100$$

$$= \frac{2.2198 + 2.1630}{2} \times 100$$

$$= 4.3828 \times 50 = 219.14$$

(iv) Marshall-Edgeworth Method: In this method, the sums of base year and current year quantities are taken as weights. The formula for constructing the index is:

$$P_{01} = \frac{\sum P_1(q_0 + q_1)}{\sum P_0(q_0 + q_1)} \times 100$$

or $$P_{01} = \frac{\sum P_1q_0 + \sum P_1q_1}{\sum P_0q_0 + \sum P_0q_1} \times 100$$

Example 10.6: For the data given in Example 10.5, compute index number of prices for 1976 with 1970 as base using the Marshall-Edgeworth formula:

Solution: Computation of price index by Marshall-Edgeworth formula:

Item	P_0	q_0	P_1	q_1	P_0q_0	P_0q_1	P_1q_0	P_1q_1
1	2	20	5	15	40	30	100	75
2	4	4	8	5	16	20	32	40
3	1	10	2	12	10	12	20	24
4	5	5	10	6	25	30	50	60
					$\Sigma P_0q_0 = 91$	$\Sigma P_0q_1 = 92$	$\Sigma P_1q_0 = 202$	$\Sigma P_1q_1 = 199$

According to Marshall-Edgeworth Formula:

$$P_{01} = \frac{\sum P_1(q_0 + q_1)}{\sum P_0(q_0 + q_1)} \times 100 = \frac{\sum P_0q_0 + \sum P_1q_1}{\sum P_0q_0 + \sum P_0q_1} \times 100$$

$$= \frac{202+199}{91+92} \times 100 = \frac{401}{183} \times 100$$

$$= 219.125$$

NOTES

(v) Kelly's Method: In this method, neither base year nor current year quantities are taken as weights. Instead, the quantities of some reference year or the average quantity of two or more years may be taken as weights. The formula for constructing the index is:

$$P_{01} = \frac{\sum P_1 q}{\sum P_0 q} \times 100$$

Where q is the quantity of some reference year.

Example 10.7: Calculate the index number of prices for 1981 with 1980 as base year for the following data, using the Kelly's method.

Item	*Quantity*	*Price in 1980*	*Price in 1981*
Bricks	10 units	100	160
Timber	7 ,,	200	210
Board	15 ,,	50	60
Sand	9 ,,	20	30
Cement	10 ,,	10	14

Solution: Computation of price index by Kelly's method:

Item	q	P_0	P_1	P_0q	P_1q
Bricks	10	100	160	1000	1600
Timber	7	200	210	1400	1470
Boards	15	50	60	750	900
Sand	9	20	30	180	270
Cement	10	10	14	100	140
				$\sum P_0 q =$ 3430	$\sum P_1 q =$ 4380

According to Kelly's method:

$$P_{01} = \frac{\sum P_1 q}{\sum P_0 q} \times 100$$

$$= \frac{4380}{3430} \times 100 = 127.697$$

(vi) Fisher's Ideal Index: This method is the geometric mean of Laspeyre's and Paasche's indices.

The formula for constructing the index is:

$$P_{01} = \sqrt{\frac{\sum P_1 q_0}{\sum P_0 q_0} \times \frac{\sum P_1 q_1}{\sum P_0 q_1}} \times 100$$

NOTES

Fisher's formula is known as ideal index because of the following reasons:

(i) It takes into account prices and quantities of both the current year as well as the base year.

(ii) It uses geometric mean which, theoretically, is the best average for constructing index numbers.

(iii) It satisfies both the time reversal test and the factor reversal test.

(iv) It is free from bias. The weight biases embodied in Laspeyre's and Paasche's methods are crossed geometrically, and thus, eliminated completely.

Example 10.8: Construct the index number of prices for the year 1980 with 1979 as base using the Fisher's Ideal Method.

	1979		*1980*	
Commodity	*Price*	*Quantity*	*Price*	*Quantity*
A	20	8	40	6
B	50	10	60	5
C	40	15	50	10
D	20	20	20	15

Solution: Construction of price index by Fisher's Ideal Formula:

Commodity	P_0	q_0	P_1	q_1	P_0q_0	P_0q_1	P_1q_0	P_1q_1
A	20	8	40	6	160	120	320	240
B	50	10	60	5	500	250	600	300
C	40	15	50	10	600	400	750	500
D	20	20	20	15	400	300	400	300
					$\Sigma P_0q_0 = 1660$	$\Sigma P_1q_1 = 1070$	$\Sigma P_1q_0 = 2070$	$\Sigma P_1q_1 = 1340$

Price index by Fisher's Ideal Formula is:

$$P_{01} = \sqrt{\frac{\sum P_1q_0}{\sum P_0q_0} \times \frac{\sum P_1q_1}{\sum P_0q_1}} \times 100$$

$$= \sqrt{\frac{2070}{1660} \times \frac{1340}{1070}} \times 100$$

$$= \sqrt{1.247 \times 1.252} \times 100 = \sqrt{1.5612} \times 100$$

$$= 1.25 \times 100 = 125$$

2. Weighted Average of Price Relatives

This method is similar to the simple average of price relatives method with the fundamental difference that explicit weights are assigned to each commodity included in the index. Since price relatives are in percentages, the weights used are value weights.

NOTES

The following steps are taken in the construction of weighted average of price relatives index:

(i) Calculate the price relatives, $\left(\frac{P_1}{P_0}\times 100\right)$, for each commodity.

(ii) Determine the value weight of each commodity in the group by multiplying its price in base year by its quantity in the base year, i.e., calculate P_0q_0 for each commodity. If, however, current year quantities are given, then the weights shall be represented by P_1q_1.

(iii) Multiply the price relative of each commodity by its value weight as calculated in (ii).

(iv) Sum up the products obtained under (iii).

(v) Divide the total (iv) above by the total of the value weights. Symbolically, index number obtained by the method of weighted average of price relatives is:

$$P_{01} = \frac{\sum\left[\left(\frac{P_1}{P_0}\times 100\right)P_0q_0\right]}{\sum P_0q_0} \text{ or } \frac{\sum PV}{\sum V}$$

The method is also known as Family Budget method.

Example 10.9: Calculate consumer price index using weighted average of price relatives method for the year 1986 with 1985 as base for the following data:

Price (in `)

Commodity	*Quantity*	*1985*	*1986*
A	100	8	12
B	25	6	8
C	10	5	15
D	20	10	25

Solution: Calculation of Consumer Price Index

Commodity	q_0	P_0	P_1	Price Relative $\left(\frac{P_1}{P_0}\times 100\right)$ or P	P_0q_0 or V	PV
A	100	8	12	150.00	800	120000
B	25	6	8	133.33	150	20000
C	10	5	15	300.00	50	15000
D	20	10	25	250.00	200	50000
					$\sum V = 1200$	$\sum PV = 205000$

NOTES

Weighted average of price relative index

or consumer price index

$$= \frac{\sum \left(\frac{P_1}{P_0} \times 100 \right) P_0 q_0}{\sum P_0 q_0} = \frac{\sum PV}{\sum V}$$

$$= \frac{205000}{1200} = 170.83$$

10.3.1 Problems in the Construction of Index Numbers

Different problems are faced in the construction of different types of index numbers. We shall deal here with only those problems which must be tackled before constructing index numbers of prices.

Definition of Purpose

It is absolutely necessary that the purpose of the index numbers be rigourously defined. This would help in deciding the nature of data to be collected, the choice of the base year, the formula to be used and other related matters. For instance, if an index number is intended to measure consumer prices, it must not include wholesale prices. Similarly, if a consumer price index number is intended to measure the changes in the cost of living of families with low incomes, great care should be exercised not to include goods ordinarily used by middle-income and upper-income groups. In fact, before constructing index numbers, we must precisely know what we want to measure, and what we intend to use this measurement for.

Selection of a Base Period

In order to make comparison between prices referring to several time periods, some point of reference is almost always established. This point of reference is called the base period. The prices of a certain time period are taken as the standard, and to them is assigned the value of 100 per cent. Though the selection of the base period would primarily depend upon the purpose of the index, the following are two important guidelines to consider in choosing a base:

(i) The base period should be a period of normal and stable economic conditions. It should be free from abnormalities and random or irregular fluctuations like wars, earthquakes, famines, strikes, lock-outs, booms, depressions, etc. Sometimes, it is difficult to choose just one year which is normal in all respects. In such cases, we can take an average of a few years as base. The process of averaging will reduce the effect of extremes.

(ii) The base year should not be too distant in the past. Since the index numbers are useful in decision-making, and economic practices are often a matter of the short run, we should choose a base which is relatively close to the year being studied. If the base year is too far in the past, we cannot make valid, meaningful comparisons since there might have been appreciable change in the tastes, customs, habits and fashion of the people during the intervening period. This would have affected the consumption pattern of the various commodities to a marked extent making comparison difficult.

Fixed Base and Chain Base: While selecting the base year, a decision has to be made whether the base shall remain fixed or not. If the period of comparison is fixed for all

NOTES

current years, it is called fixed base method. If, on the other hand, the prices of the current year are linked with the prices of the preceding year and not with the fixed year or period, it is called chain base method. Chain base method is useful in cases where there are quick and frequent changes in fashion, tastes and habits of the people. In such cases comparison with the preceding year is more worthwhile.

Selection of Commodities or Items

While constructing an index number, it is not possible to take into account all the items whose price changes are to be represented by the index number. Hence, the need for selecting a sample. For instance, while constructing a general purpose wholesale price index, it is impossible to take all the items. Thus, only a few representative items are selected from the whole lot. While selecting the sample the following points should be kept in mind:

(i) The selected commodity or item should be representative of the tastes, customs and necessities of the people to whom the index number relates.

(ii) It should be stable in quality and as far as possible should be standardized or graded so that it can easily be identified after a time lapse.

(iii) The sample should be as large as possible. Theoretically, the larger the number of items, the more accurate would be the results disclosed by an index number. But it must be noted that larger the number of items, the greater shall be the cost and time taken. Therefore, the number of items should be determined on the basis of the purpose of the index as well as the basis of funds available and the time within which the index numbers must be ready.

(iv) As different varieties of a commodity are sold in the market, a decision has to be made as to which variety should be included in the index numbers. Ordinarily, all those varieties which are in common use should be included, and to avoid extra weightage for any commodity, the prices of these varieties should be averaged before their inclusion in the index numbers.

Obtaining Price Quotations

After selecting the items, the next problem is to collect their prices. The price of a commodity varies from place to place and even from shop to shop in the same market. Just as it is not possible to include all the commodities in an index number, similarly it is impracticable to collect price quotations from all places where a commodity is bought or sold. Thus, a selection is to be made of representative places and shops. Generally, such places and shops are selected where the commodity is bought and sold in large quantities. After selecting the places and shops from where price quotations are to be obtained, the next step is to appoint some representatives who will supply the price quotations from time to time. While appointing such representatives, it must be ensured that they are unbiased and are reliable. If price quotations are published by some reliable agency, journal or magazine, then such price quotations may also be used.

Since prices can be quoted in two ways, i.e., either by expressing the quantity of commodity per unit of money or by expressing the quantity of money per unit of commodity, a decision has to be made regarding the manner in which prices are to be quoted. The second method, that of quoting price per unit of commodity, is free from confusion and is generally adopted. Thus it is better to quote the price of a commodity X as 50 paise per kg rather than quoting it as 2 kg per one rupee.

NOTES

Another decision in regard to price quotations is whether the wholesale prices or the retail prices are to be collected. This will depend upon the purpose of the index. For instance, in the case of consumer price index, the wholesale prices are not representative at all because consumers do not generally make their purchases in bulk from the wholesale market. Similarly, if the prices of certain commodities are controlled by the government, then such controlled prices should be taken into account and not the black market prices which may be much higher.

Another thing associated with the price quotations is the decision with regard to the number of price quotations to be collected every week or every month. In general, the larger the number of quotations, the better it is. Ordinarily, however, at least one quotation per week in case of weekly indices, and at least four quotations per month in case of monthly indices are essential. In deciding the frequency of price quotations, the guiding principle is that the number of quotations should be such that the agency supplying the quotations can easily and regularly send them.

Choice of Average

Since index numbers are specialized averages, a decision has to be made as to which particular average, i.e., arithmetic mean, mode, median, harmonic mean or geometric mean should be used for the construction of index numbers. Mode, median and harmonic mean are almost never used in the construction of index numbers.

Therefore, a choice has to be made between arithmetic mean and geometric mean. Though theoretically geometric mean is better for the purpose, arithmetic mean due to its simplicity of computation is more commonly used.

Choice of Weights

All items included in the index numbers are not of equal importance, and thus it is necessary that some suitable method is devised by which the varying importance of different items is taken into account. This is done by assigning 'weights'. The term 'weight' refers to the relative importance of different items in the construction of index.

There are two methods of assigning weights: (i) Implicit, and (ii) Explicit. In the case of implicit weighting, a commodity or its variety is included in the index a number of times. In the case of explicit weighting, on the other hand, some outward evidence of importance of various items in the index is given. Explicit weights are of two types: (i) Quantity weights, and (ii) Value weights. A quantity weight means the amount of commodity produced, consumed or distributed in a particular time period. The quantity weights are used when the aggregative method of constructing index numbers is used. On the other hand, if the average of price relatives method is used, then values are used as weights.

Selection of an Appropriate Formula

A large number of formulae have been devised for constructing the index numbers. A decision has, therefore, to be made as to which formula is the most suitable for the purpose. The choice of the formula depends upon the availability of the data regarding the prices and quantities of the selected commodities in the base and/or current year.

Quantity or Volume Index Numbers

Price indices measure changes in the price level of certain commodities. On the other hand, quantity or volume index numbers measure the changes in the physical volume of goods produced, distributed or consumed. These indices are important indicators of the level of output in the economy or in parts of it.

NOTES

In constructing quantity index numbers, the problems facing the statistician are similar to the ones faced by him in constructing price indices. In this case we measure changes in quantities, and when we weigh, we use prices as weights.

The quantity indices can be obtained easily by replacing p by q and vice versa in the various formulae discussed earlier.

The quantity index by different methods is as follows:

(i) Laspeyre's method: $$Q_{01} = \frac{\Sigma q_1 P_0}{\Sigma q_0 P_0} \times 100$$

(ii) Paasche's method: $$Q_{01} = \frac{\Sigma q_1 P_1}{\Sigma q_0 P_1} \times 100$$

(iii) Bowley-Drobisch Method: $$Q_{01} = \frac{\frac{\Sigma q_1 P_0}{\Sigma q_0 P_0} + \frac{\Sigma q_1 P_1}{\Sigma q_0 P_1}}{2} \times 100$$

(iv) Marshall-Edgeworth Method: $$Q_{01} = \frac{\Sigma q_1 (P_0 + P_1)}{\Sigma q_0 (P_0 + P_1)} \times 100$$

(v) Fisher's Ideal Index: $$Q_{01} \sqrt{\frac{\Sigma q_1 P_0}{\Sigma q_0 P_0} \times \frac{\Sigma q_1 P_1}{\Sigma q_0 P_1}} \times 100$$

(vi) Kelly's method: $$Q_{01} = \frac{\Sigma q_1 P}{\Sigma q_0 P} \times 100$$

Example 10.10: Compute quantity index for the year 1982 with base 1980 = 100, for the following data, using (i) Laspeyre's method (ii) Paasche's method, (iii) Bowley-Drobisch method, (iv) Marshall-Edgeworth method, and (v) Fisher's ideal formula.

	Prices		*Quantities*	
Commodity	*1980*	*1982*	*1980*	*1982*
A	5.00	6.50	5	7
B	7.75	8.80	6	10
C	9.63	7.75	4	6
D	12.50	12.75	9	9

Solution: Computation of quantity index

Commodity	P_0	q_0	P_1	q_1	q_0P_0	q_0P_1	q_1P_0	q_1P_1
A	5.00	5	6.50	7	25.00	32.50	35.00	45.50
B	7.75	6	8.80	10	46.50	52.80	77.50	88.00
C	9.63	4	7.75	6	38.52	31.00	57.78	46.50
D	12.50	9	12.75	9	112.50	114.75	112.50	114.75
					$\Sigma q_0P_0 = 222.52$	$\Sigma q_0P_1 = 231.05$	$\Sigma q_1P_0 = 282.78$	$\Sigma q_1P_1 = 294.75$

NOTES

(i) Laspeyre's quantity index or $Q_{01} = \frac{\Sigma q_1 P_0}{\Sigma q_0 P_0} \times 100$

$$= \frac{282.78}{222.52} \times 100 = 127.08$$

(ii) Paasche's quantity index or $Q_{01} = \frac{\Sigma q_1 P_1}{\Sigma q_0 P_1} \times 100$

$$= \frac{294.75}{231.05} \times 100 = 127.57$$

(iii) Bowley-Drobisch quantity index or $Q_{01} = \frac{\frac{\Sigma q_1 P_0}{\Sigma q_0 P_0} + \frac{\Sigma q_1 P_1}{\Sigma q_0 P_1}}{2} \times 100$

$$= \frac{\frac{282.78}{222.52} + \frac{294.75}{231.05}}{2} \times 100$$

$$= \frac{1.2708 + 1.2757}{2} \times 100$$

$$= 127.325$$

(iv) Marshall-Edgeworth quantity index or $Q_{01} = \frac{\Sigma q_1 P_0 + \Sigma q_1 P_1}{\Sigma q_0 P_0 + \Sigma q_0 P_1} \times 100$

$$= \frac{282.78 + 294.75}{222.52 + 231.05} \times 100$$

$$= 127.329$$

(v) Quantity index by Fisher's ideal formula or Q_{01}

$$= \sqrt{\frac{\Sigma q_1 P_0}{\Sigma q_0 P_0} \times \frac{\Sigma q_1 P_1}{\Sigma q_0 P_1}} \times 100$$

$$= \sqrt{\frac{282.78}{222.52} \times \frac{294.75}{231.05}} \times 100$$

$$= 1.273 \times 100$$

$$= 127.3$$

Value Index Numbers

Value means price times quantity. Thus, a value index V is the sum of the value of a given year divided by the sum of the values for the base year. The formula, therefore, is:

$$V = \frac{\Sigma P_1 q_1}{\Sigma P_0 q_0} \times 100$$ where V = value index

Since in most cases, the value figures given in the formula may be stated more simply as:

$$V = \frac{\sum V_1}{\sum V_0}$$

In this type of index, both price and quantity are variable in the numerator. Weights are not to be applied because they are inherent in the value figures. A value index, therefore, is an aggregate of values.

Tests of Consistency

As there are several formulae for constructing index numbers, the problem is to select the most appropriate formula in a given situation. Prof. Irving Fisher has suggested two tests for selecting an appropriate formula. These are as follows:

1. Time reversal test
2. Factor reversal test

Time Reversal Test

According to Prof. Fisher, the formula for calculating the index should be such that it gives the same ratio between one point of comparison and another, no matter which of the two is taken as base. In other words, the index number prepared forward should be the reciprocal of the index number prepared backward. Thus, if from 1982 to 1983, the prices of a basket of goods have increased from ` 400 to ` 800, so that the index number for 1983 with 1982 as base is 200 per cent. Now if the index number for 1983 with 1982 as base is 200 per cent, the index number for 1982 with 1983 with base should be 50 per cent. One figure is reciprocal of the other and their product (2 × 0.5) is unity. Therefore, time reversal test is satisfied if $P_{01} \times P_{10} = 1$.

Time reversal test is satisfied by:

(i) Fisher's Ideal Formula

(ii) Marshall-Edgeworth Method

(iii) Kelly's Method

(iv) Simple Geometric mean of Price relatives

Factor Reversal Test

According to Prof. Fisher, the formula for constructing the index number should permit not only the interchange of the two times without giving inconsistent results, it should also permit the interchange of weights without giving inconsistent results.

Simply stated, the test is satisfied if the change in price multiplied by the change in quantity is equal to the total change in value. Thus, factor reversal test is satisfied if:

$$P_{01} \times Q_{01} = \frac{\sum P_1 q_1}{\sum P_0 q_0}$$

Where, P_{01} represents change in price in the current year, Q_{01} represents change in quantity in the current year, $\Sigma P_1 q_1$ represents total value in the current year, and $\Sigma P_0 q_0$ represents total value in the base year.

The factor reversal test is satisfied only by Fisher's Ideal Formula. Thus, Fisher's formula satisfies both time reversal test and factor reversal test.

NOTES

Proof

According to Fisher's Ideal Index:

$$P_{01} = \sqrt{\frac{\sum P_1 q_0}{\sum P_0 q_0} \times \frac{\sum P_1 q_1}{\sum P_0 q_1}}$$

$$P_{10} = \sqrt{\frac{\sum P_0 q_1}{\sum P_1 q_1} \times \frac{\sum P_0 q_0}{\sum P_1 q_0}}$$

$$Q_{01} = \sqrt{\frac{\sum q_1 P_0}{\sum q_0 P_0} \times \frac{\sum q_1 P_1}{\sum q_0 P_1}}$$

(i) Thus $$P_{01} \times P_{10} = \sqrt{\frac{\sum P_1 q_0}{\sum P_0 q_0} \times \frac{\sum P_1 q_1}{\sum P_0 q_1} \times \frac{\sum P_0 q_1}{\sum P_1 q_1} \times \frac{\sum P_0 q_0}{\sum P_1 q_0}} = \sqrt{1} = 1$$

Hence, the time reversal test is satisfied.

(ii) Similarly, according to Fisher's Ideal Formula:

$$P_{01} \times Q_{01} = \sqrt{\frac{\Sigma P_1 q_0}{\Sigma P_0 q_0} \times \frac{\Sigma P_1 q_1}{\Sigma P_0 q_1} \times \frac{\Sigma q_1 P_0}{\Sigma q_0 P_0} \times \frac{\Sigma q_1 P_1}{\Sigma q_0 P_1}}$$

$$= \sqrt{\frac{\Sigma P_1 q_1 \times \Sigma q_1 P_1}{\Sigma P_0 q_0 \times \Sigma q_0 P_0}} = \frac{\Sigma P_1 q_1}{\Sigma P_0 q_0}$$

Hence, the factor reversal test is also satisfied by Fisher's Ideal Formula.

Besides these two tests, two other tests have been suggested by some authors.

These are: 1. Unit test 2. Circular test

Unit Test

According to unit test, the formula for constructing index numbers should be independent of the units in which prices and quantities are quoted. This test is satisfied only by simple aggregative index method.

Circular Test

This test is just an extension of the time reversal test for more than two periods and is based on the shiftability of the base period. This test requires the index number to work in a circular manner such that if an index is constructed for the year *a* on base year *b*, and for the year *b* on base year *c*, we should get the same result as if we calculate directly an index for year *a* on base year *c* without going through *b* as an intermediary. Thus, if there are three periods *a*, *b* and *c*, the circular test is satisfied if,

$$P_{01} \times P_{12} \times P_{10} = 1$$

The circular test is satisfied only by the index number formula based on:

(i) Simple aggregate of prices.

(ii) Kelly's method or fixed weighted aggregate of prices.

An index which satisfies this list has the advantage of reducing the computations every time a change in the base year has to be made. Such indices can be adjusted from year to year without referring each time to the original base.

Example 10.11: From the following data, show that Fisher's Ideal Index satisfies both following time reversal test and factor reversal test.

	1980		*1981*	
Commodity	*Price*	*Quantity*	*Price*	*Quantity*
A	4	10	5	8
B	6	8	9	7
C	14	5	7	12
D	3	12	6	8
E	5	7	8	5

Solution: Computation for time reversal test and factor reversal test

Commodity	P_0	q_0	P_1	q_1	P_0q_0	P_0q_1	P_1q_0	P_1q_1
A	4	10	5	8	40	32	50	40
B	6	8	9	7	48	42	72	63
C	14	5	7	12	70	168	35	84
D	3	12	6	8	36	24	72	48
E	5	7	8	5	35	25	56	40
							$\sum P_1q_0 = 285$	$\sum P_1q_1 = 275$

$\sum P_0q_0 = 229$ $\quad \sum P_0q_1 = 291$

(i) Time reversal test is satisfied when $P_{01} \times P_{10} = 1$.

According to Fisher's ideal index:

$$P_{01} = \sqrt{\frac{\sum P_1q_0}{\sum P_0q_0} \times \frac{\sum P_1q_1}{\sum P_0q_1}} \text{ and } P_{10} = \sqrt{\frac{\sum P_0q_1}{\sum P_1q_1} \times \frac{\sum P_0q_0}{\sum P_1q_0}}$$

$$P_{01} \times P_{01} = \sqrt{\frac{285}{229} \times \frac{275}{291} \times \frac{291}{275} \times \frac{229}{285}} = \sqrt{1} = 1$$

Hence, time reversal test is satisfied.

(ii) Factor reversal test is satisfied when $P_{01} \times Q_{01} = \frac{\sum P_1q_1}{\sum P_0q_0}$.

$$P_{01} = \sqrt{\frac{\sum P_1q_0}{\sum P_0q_0} \times \frac{\sum P_1q_1}{\sum P_0q_1}} \text{ and } Q_{01} = \sqrt{\frac{\sum q_1P_0}{\sum q_0P_0} \times \frac{\sum q_1P_1}{\sum q_0P_1}}$$

$$P_{01} \times Q_{01} = \sqrt{\frac{285}{229} \times \frac{275}{291} \times \frac{291}{229} \times \frac{275}{285}} = \sqrt{\frac{275 \times 275}{229 \times 229}}$$

$$= \frac{275}{229} = \frac{\sum P_1q_1}{\sum P_0q_0}$$

Hence, the factor reversal test is satisfied.

NOTES

Fixed and Chain Base Indices

As stated earlier, the base may be fixed or changing. It is said to be fixed when the period of comparison or the base year is fixed for all current years. Thus, if the indices of 1971, 1972, 1973 and 1974 are all calculated with 1970 as the base year, such indices will be called fixed base indices. If, on the other hand, the whole series of index numbers is not related to any one base period, but the indices for different years are obtained by relating each year's price to that of the immediately preceding year, the indices so obtained are called chain base indices. For instance, in the case of chain base indices, for 1974, 1973 will be the base; for 1973, 1972 will be the base; for 1972, 1971 will be the base, and so on. The relatives obtained by the chain base method are called link relatives, whereas the relatives obtained by the fixed base method are called chain relatives.

Example 10.12: From the following data relating to the wholesale prices of wheat for six years, construct index numbers using (a) 1980 as base, and (b) by chain base method.

Year	*Price (per quintal)* `	*Year*	*Price (per quintal)* `
1980	100	1983	130
1981	120	1984	140
1982	125	1985	150

Solution: (a) Computation of index numbers with 1980 as base:

Year	*Price of wheat*	*Index Number (1980 = 100)*	*Year*	*Price of Wheat*	*Index No. (1980 = 100)*
1980	100	100	1983	130	$\frac{130}{100}\times100=130$
1981	120	$\frac{120}{100}\times100=120$	1984	140	$\frac{140}{100}\times100=140$
1982	125	$\frac{125}{100}\times100=125$	1985	150	$\frac{150}{100}\times100=150$

(b) Construction of link relative indices (chain base method)

Year	*Price of Wheat*	*Link Relative Index*	*Year*	*Price of Wheat*	*Link Relative Index*
1980	100	100	1983	130	$\frac{130}{125}\times100=104$
1981	120	$\frac{120}{100}\times100=120$	1984	140	$\frac{140}{130}\times100=107.692$
1982	125	$\frac{125}{120}\times100=104.167$	1985	150	$\frac{150}{140}\times100=107.14$

NOTES

Conversion of Link Relatives into Chain Relatives

Chain relatives or chain indices can be obtained either directly or by converting link relatives into chain relatives with the help of the following formula:

$$\text{Chain relative for current year} = \frac{\text{Link relative for the current year} \times \text{Chain relative for the previous year}}{100}$$

Taking the data from Example 10.12, we can show the method of conversion as follows:

Year	*Price of wheat*	*Link relative*	*Chain relative*
1980	100	100.00	100
1981	120	120.00	
1982	125	104.167	$\frac{104.167 \times 120}{100} = 125$
1983	130	104.00	$\frac{104 \times 125}{100} = 130$
1984	140	107.692	$\frac{107.692 \times 130}{100} = 140$
1985	150	107.14	$\frac{107.14 \times 140}{100} = 150$

$\frac{120 \times 100}{100} = 120$

Base Shifting

Sometimes it becomes necessary to shift the base from one period to another. This becomes necessary either because the previous base has become too old and useless for comparison purposes or because comparison has to be made with another series of index numbers having different base period. This can be done in following two ways:

(i) By reconstructing the series with the new base. This means that the relatives of each individual item are constructed with the new base and thus an entirely new series is formed.

(ii) By using a shorter method which is as follows: divide each index number of the series by the index number of the time period selected as new base and multiply the quotient by 100. Symbolically,

$$\text{Index Number (based on new base year)} = \frac{\text{Current year's old index number}}{\text{New base year's old index number}} \quad 100$$

Example 10.13: The following are the index numbers of pries with 1939 as base:

Year:	1939	1940	1945	1950	1955	1960
Index Number:	100	110	120	200	400	380

Shift the base to the year 1950.

NOTES

Solution: Index numbers with 1950 as base (1950-100)

Year	*Index (1939 = 100)*	*Index Number (1950 = 100)*
1939	100	$\frac{100}{200} \times 100 = 50$
1940	110	
1945	120	
1950	200	
1955	400	$\frac{400}{200} \times 100 = 200$
1960	380	$\frac{380}{200} \times 100 = 190$

Splicing

Sometimes, an index number series is discontinued because its base has become too old and so it has lost its utility. A new series of index numbers may be computed with some recent year as base. For instance, the weights of an index number may have become out of date and a new index with new weights may be constructed. This would result in two series of index numbers. It may sometimes be necessary to connect the two series of index number into one continuous series. The procedure employed for connecting an old series of index numbers with a revised series in order to make the series continuous is called splicing. The process of splicing is very simple and is similar to the one used in shifting the base. The spliced index numbers are calculated with the help of the following formula:

New Base Year's

$$\text{Spliced Index Number} = \frac{\text{Current Year's New Index Number} \times \text{Old Index No.}}{100}$$

Example 10.14: Index *A* was started in 1969 and continued upto 1975 in which year another index *B* was started. Splice the index *B* to index *A* so that a continuous series of index numbers from 1969 upto date may be available:

	Year	1969	1970	1971	1972	1973	1974	1975
(A)	Index Numbers (Old)	100	120	130	200	300	350	400
	Year	1975	1976	1977	1978	1979	1980	
(B)	Index Numbers (New)	100	110	90	110	98	96	

Solution: Index *B* spliced to Index *A*

Year	*Old Index Nos.*	*New Index Nos.*	*Index B Spliced to Index A (Base 1969 = 100)*
1969	100		
1970	120		
1971	130		
1972	200		
1973	300		
1974	350		
1975	400	100	$\frac{400\times100}{100}=400$
1976		110	$\frac{400\times110}{100}=440$
1977		90	$\frac{400\times90}{100}=360$
1978		110	$\frac{400\times110}{100}=440$
1979		98	$\frac{400\times98}{100}=392$
1980		96	$\frac{400\times96}{100}=384$

NOTES

Splicing is very useful for making comparison between new and old index numbers.

Deflating

Deflating is the process of making allowances for the effect of changing price levels. With increasing price levels, the purchasing power of money is reduced. As a result, the real wage figures are reduced and the real wages become less than the money wages. To get the real wage figure, the money wage figure may be reduced to the extent the price level has risen. The process of calculating the real wages by applying index numbers to the money wages so as to allow for the change in the price level is called deflating. Thus, deflating is the process by which a series of money wages or incomes can be corrected for price changes to find out the level of real wages or incomes. This is done with the help of the following formula:

$$\text{Real wage} = \frac{\text{Money wage}}{\text{Price index}}\times100$$

$$\text{Real wage index} = \frac{\text{Real wages for the current year}}{\text{Real wages for the base year}}\times100$$

NOTES

Example 10.15: The average of monthly wages average in different years is as follows:

Year	:	1977	1978	1979	1980	1981	1982	1983
Wages (`)	:	200	240	350	360	360	380	400
Price Index	:	100	150	200	220	230	250	250

Calculate real wages index numbers.

Solution: Construction of real wage indices

Year	*Wages (`)*	*Price index*	*Real wages*	*Real wages index (1977 = 100)*
1977	200	100	$\frac{200}{100} \times 100 = 200$	100
1978	240	150	$\frac{240}{150} \times 100 = 160$	$\frac{160}{200} \times 100 = 80$
1979	350	200		$\frac{175}{200} \times 100 = 87.5$
1980	360	220	$\frac{360}{220} \times 100 = 163.63$	$\frac{163.63}{200} \times 100 = 81.81$
1981	360	230	$\frac{360}{230} \times 100 = 156.52$	$\frac{156.52}{200} \times 100 = 78.26$
1982	380	250	$\frac{380}{250} \times 100 = 152$	$\frac{152}{200} \times 100 = 76$
1983	400	250		$\frac{160}{200} \times 100 = 80$

5.4 PRICE INDEX AND COST OF LIVING INDEX

A cost of living index is a theoretical price index that measures relative cost of living over time or regions. It is an index that measures differences in the price of goods and services, and allows for substitutions to other items as prices vary.

There are many different methodologies that have been developed to approximate cost of living indexes, including methods that allow for substitution among items as relative prices change. The following examples will make the concept clear.

Problem 1: From the following data, construct index number of prices for 1986 with 1980 as base, using (i) Laspeyre's method, (ii) Paasche's method, (iii) Bowley-Drobisch method, (iv) Marshall-Edgeworth method, (v) Fisher's ideal formula.

NOTES

	1980		1986	
Commodity	*Price Per Unit*	*Expenditure in Rupees*	*Price Per Unit*	*Expenditure in Rupees*
A	2	10	4	16
B	3	12	6	18
C	1	8	2	14
D	4	20	8	32

Solution: Since we are given the price and the total expenditure for the year 1980 and 1986, we shall first calculate the quantities for the two years by dividing the expenditure by price, and then we shall calculate the index numbers as follows:

Commodity	P_0	q_0	P_1	q_1	P_0q_0	P_0q_1	P_1q_0	P_1q_1
A	2	5	4	4	10	8	20	16
B	3	4	6	3	12	9	24	18
C	1	8	2	7	8	7	16	14
D	4	5	8	4	20	16	40	32
						$\sum P_0q_1 = 40$	$\sum P_1q_0 = 100$	$\sum P_1q_1 = 80$

$\sum P_0q_0 = 50$

(i) Laspeyre's price index or $P_{01} = \dfrac{\sum P_1q_0}{\sum P_0q_0} \times 100$

$$= \frac{100}{50} \times 100 = 200$$

(ii) Paasche's price index or $P_{01} = \dfrac{\sum P_1q_1}{\sum P_0q_1} \times 100$

$$= \frac{80}{40} \times 100 = 200$$

(iii) Bowley-Drobisch price index or $P_{01} = \dfrac{\dfrac{\sum P_1q_0}{\sum P_0q_0} + \dfrac{\sum P_1q_1}{\sum P_0q_1}}{2} \times 100$

$$= \frac{\dfrac{100}{50} + \dfrac{80}{40}}{2} \times 100 = 200$$

(iv) Marshall-Edgeworth price index or $P_{01} = \dfrac{\sum p_1q_0 + \sum p_1q_1}{\sum p_0q_0 + \sum p_0q_1} \times 100$

$$= \frac{100 + 80}{50 + 40} \times 100$$

$$= 200$$

NOTES

(v) Fisher's Ideal index of price or $P_{01}\sqrt{\frac{\sum p_1q_0}{\sum p_0q_0}\times\frac{\sum p_1q_1}{\sum p_0q_1}}\times 100$

$$=\sqrt{\frac{100}{50}\times\frac{80}{40}}\times 100$$

$$=\sqrt{2\times 2}\times 100$$
$$=200$$

Problem 2: From the following data construct index number of quantities, and of prices for 1970 with 1966 as base using (i) Laspeyre's formula, (ii) Paasche's formula, and (iii) Fisher's Ideal formula.

	1996		*1970*	
Commodity	*Price* `	*Quantity (Units)*	*Price* `	*Quantity (Units)*
A	5.20	100	6	150
B	4.00	80	5	100
C	2.50	60	5	72
D	12.00	30	9	33

Solution: Calculation of quantity index and price index by (i) Laspeyre's formula, (ii) Paasche's formula, and (iii) Fisher's Ideal formula.

Commodity	P_o	q_0	P_1	q_1	P_0q_0	P_0q_1	P_1q_0	P_1q_1
A	5.20	100	6	150	520	780	600	900
B	4.00	80	5	100	320	400	400	500
C	2.50	60	5	72	150	180	300	360
D	12.00	30	9	33	360	396	270	297
					$\sum P_0q_0 = 1350$	$\sum P_0q_1 = 1756$	$\sum P_1q_0 = 1570$	$\sum P_1q_1 = 2057$

A. Quantity index number for 1970 with 1966 as base by:

(i) Quantity index by Laspeyre's formula or $Q_{01}=\frac{\sum q_1p_0}{\sum q_0p_0}\times 100$

$$=\frac{1756}{1350}\times 100=130.07$$

(ii) Quantity index by Paasche's formula or $Q_{01}=\frac{\sum q_1p_1}{\sum q_0p_1}\times 100$

$$=\frac{2057}{1570}\times 100=131.02$$

(iii) Quantity index by Fisher's ideal formula or Q_{01}

$$=\sqrt{\frac{\sum q_1P_0}{\sum q_0P_0}\times\frac{\sum q_1P_1}{\sum q_0P_1}}\times 100$$

NOTES

$$\sqrt{\frac{1756}{1350} \quad \frac{2057}{1570}} \quad 100 \quad \sqrt{1.3007 \quad 1.3102} \quad 100$$

$$= \sqrt{1.704177} \times 100 = 130.54$$

B. Price index number for 1970 with 1966 as base by:

(i) Laspeyre's formula or P_{01} $= \dfrac{\sum P_1q_0}{\sum P_0q_0} \times 100$

$$= \frac{1570}{1350} \times 100 = 116.296$$

(ii) Paasche's formula or P_{01} $= \dfrac{\sum p_1q_1}{\sum p_0q_1} \times 100$

$$= \frac{2057}{1756} \times 100 = 117.14$$

(iii) Fisher's ideal index of price or $P_{01} = \sqrt{\dfrac{\sum p_1q_0}{\sum p_0q_0} \times \dfrac{\sum p_1q_1}{\sum p_0q_1}} \times 100$

$$\sqrt{\frac{1570}{1350} \quad \frac{2057}{1756}} \quad 100$$

$$= \sqrt{1.16296 \times 1.17141} \times 100$$

$$= \sqrt{1.362302} \quad 100$$

$$= 1.167 \times 100 = 116.7$$

$\sum P_0q_0$
$= 9942$

Problem 3: From the following data, calculate the price index by Fisher's ideal formula and then verify that Fisher's ideal formula satisfies both time reversal test and factor reversal test.

Commodity	*Base year*		*Current year*	
	Price (`)	*Quantity* (,000 tonnes)	*Price* (`)	*Quantity* (,000 tonnes)
A	56	71	50	26
B	32	107	30	83
C	41	62	28	48

Solution: Calculation of price index by Fisher's ideal formula and computations for time reversal test and factor reversal test.

Commodity	P_0	q_0	P_1	q_1	P_0q_0	P_0q_1	P_1q_0	P_1q_1
A	56	71	50	26	3976	1456	3550	1300
B	32	107	30	83	3424	2659	3210	2490
C	41	62	28	48	2542	1968	1736	1344
						$\sum P_0q_1 = 6083$	$\sum P_1q_0 = 8496$	$\sum P_1q_1 = 5134$

NOTES

(i) Fisher's ideal index of price or $P_{01} = \sqrt{\frac{\sum P_1 q_0}{\sum P_0 q_0} \times \frac{\sum P_1 q_1}{\sum P_0 q_1}} \times 100$

$$= \sqrt{\frac{8496}{9942} \times \frac{5134}{6083}} \times 100$$

$$= \sqrt{0.8544 \times 0.844} \times 100$$

$$0.8492 \times 100 = 84.92$$

(ii) Time reversal test is satisfied if $P_{01} \times p_{10} = 1$.

According to Fisher's ideal formula

$$P_{01} \times P_{10} = \sqrt{\frac{8496}{9942} \times \frac{5134}{6083} \times \frac{6083}{5134} \times \frac{9942}{8496}} = \sqrt{1} = 1$$

Hence the time reversal test is satisfied by Fisher's ideal formula.

(iii) Factor Reversal Test is satisfied if .

$$P_{01} = \sqrt{\frac{\sum P_1 q_0}{\sum P_0 q_0} \times \frac{\sum P_1 q_1}{\sum P_0 q_1}} \text{ and } Q_{01} = \sqrt{\frac{\sum q_1 P_0}{\sum q_0 P_0} \times \frac{\sum q_1 P_1}{\sum q_0 P_1}}$$

or $$P_{01} \times Q_{01} = \sqrt{\frac{8496}{9942} \times \frac{5134}{6083} \times \frac{6083}{9942} \times \frac{5134}{8496}} = \sqrt{\frac{5134}{9942} \times \frac{5134}{9942}}$$

$$= \frac{5134}{9942} = \frac{\sum P_1 q_1}{\sum P_0 q_0}$$

Hence, the factor reversal test is satisfied by Fisher's ideal formula.

Check Your Progress

1. Name the two types of methods for constructing index numbers.
2. What are the different methods for constructing weighted aggregate of price index?
3. Define the term chain base method.
4. What do you mean by the term weight?
5. What is value index?
6. Define the term splicing.
7. What do you mean by deflating?
8. What are uses of index numbers?

5.5 COMPONENTS OF TIME SERIES

The time series analysis method is quite accurate where future is expected to be similar to past. The underlying assumption in time series is that the same factors will continue to influence the future patterns of economic activity in a similar manner as in the past. These techniques are fairly sophisticated and require experts to use these methods.

The classical approach is to analyse a time series in terms of four distinct types of variations or separate components that influence a time series.

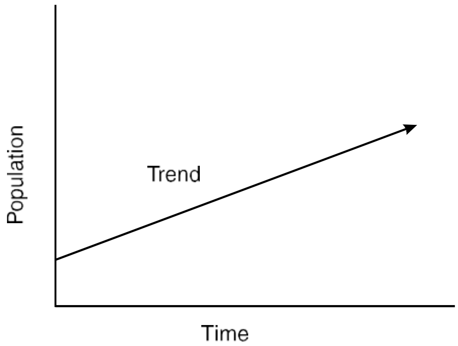
1. Secular Trend or Simply Trend (*T*)

Trend is a general long-term movement in the time series value of the variable (*Y*) over a fairly long period of time. The variable (*Y*) is the factor that we are interested in evaluating for the future. It could be sales, population, crime rate, and so on.

NOTES

Trend is a common word, popularly used in day-to-day conversation, such as population trends, inflation trends and birth rate. These variables are observed over a long period of time and any changes related to time are noted and calculated and a trend of these changes is established. There are many types of trends; the series may be increasing at a slow rate or at a fast rate or these may be decreasing at various rates. Some remain relatively constant and some reverse their trend from growth to decline or from decline to growth over a period of time. These changes occur as a result of the general tendency of the data to increase or decrease as a result of some identifiable influences.

If a trend can be determined and the rate of change can be ascertained, then tentative estimates on the same series values into the future can be made. However, such forecasts are based upon the assumption that the conditions affecting the steady growth or decline are reasonably expected to remain unchanged in the future. A change in these conditions would affect the forecasts. As an example, a time-series involving increase in population over time can be shown as,



2. Cyclical Fluctuations (*C*)

Cyclical fluctuations refer to regular swings or patterns that repeat over a long period of time. The movements are considered cyclical only if they occur after time intervals of more than one year. These are the changes that take place as a result of economic booms or depressions. These may be up or down, and are recurrent in nature and have a duration of several years—usually lasting for two to ten years. These movements also differ in intensity or amplitude and each phase of movement changes gradually into the phase that follows it. Some economists believe that the business cycle completes four phases every twelve to fifteen years. These four phases are: prosperity, recession, depression and recovery. However, there is no agreement on the nature or causes of these cycles.

Even though, measurement and prediction of cyclical variation is very important for strategic planning, the reliability of such measurements is highly questionable due to the following reasons:

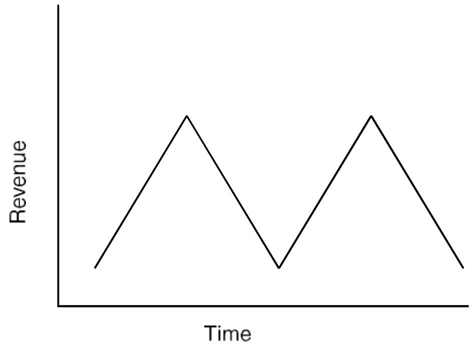
(i) These cycles do not occur at regular intervals. In the twenty-five years from 1956 to 1981 in America, it is estimated that the peaks in the cyclical activity of the overall economy occurred in August 1957, April 1960, December 1969, November 1973 and January 1980.[1] This shows that they differ widely in timing, intensity and pattern, thus making reliable evaluation of trends very difficult.

[1] Mark L. Berenson and David M. Levine, *Basic Business Statistics* (New Jersey: Prentice-Hall, 1983), 618.

NOTES

(ii) The cyclic variations are affected by many erratic, irregular and random forces which cannot be isolated and identified separately, nor can their impact be measured accurately.

The cyclic variation for revenues in an industry against time is shown graphically as follows:



3. Seasonal Variation (*S*)

Seasonal variation involves patterns of change that repeat over a period of one year or less. Then they repeat from year to year and they are brought about by fixed events. For example, sales of consumer items increase prior to Christmas due to gift giving tradition. The sale of automobiles in America are much higher during the last three to four months of the year due to the introduction of new models. This data may be measured monthly or quarterly.

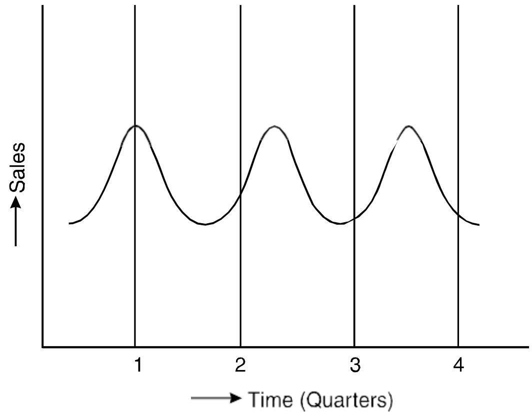
Since these variations repeat during a period of twelve months, they can be predicted fairly and accurately. Some factors that cause seasonal variations are as follows:

(i) **Season and Climate:** Changes in the climate and weather conditions have a profound effect on sales. For example, the sale of umbrellas in India is always more during monsoons. Similarly, during winter, there is a greater demand for woollen clothes and hot drinks, while during summer months there is an increase in the sales of fans and air conditioners.

(ii) **Customs and Festivals:** Customs and traditions affect the pattern of seasonal spending. For example, Mother's Day or Valentine's Day in America see increase in gift sales preceding these days. In India, festivals, such as Baisakhi and Diwali mean a big demand for sweets and candy. It is customary all over the world to give presents to children when they graduate from high school or college. Accordingly, the month of June, when most students graduate, is a time for the increase of sale for presents befitting the young.

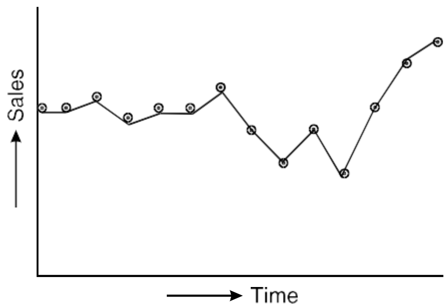
An accurate assessment of seasonal behaviour is an aid in business planning and scheduling, such as in the area of production, inventory control, personnel, advertising, and so on. The seasonal fluctuations over four repeating quarters in a given year for sale of a given item is illustrated as:

NOTES



4. Irregular or Random Variation (*I*)

These variations are accidental, random or simply due to chance factors. Thus, they are wholly unpredictable. These fluctuations may be caused by such isolated incidents as floods, famines, strikes or wars. Sudden changes in demand or a breakthrough in technological development may be included in this category. Accordingly, it is almost impossible to isolate and measure the value and the impact of these erratic movements on forecasting models or techniques. This phenomenon may be graphically shown as follows:



It is traditionally acknowledged that the value of the time series (*Y*) is a function of the impact of variable trend (*T*), seasonal variation (*S*), cyclical variation (*C*) and irregular fluctuation (*I*). These relationships may vary depending upon assumptions and purposes. The effects of these four components might be additive, multiplicative, or combination thereof in a number of ways. However, the traditional time series analysis model is characterized by multiplicative relationship, so that:

$$Y = T \times S \times C \times I$$

This model is appropriate for those situations where percentage changes best represent the movement in the series and the components are not viewed as absolute values but as relative values.

Another approach to define the relationship may be additive, so that:

$$Y = T + S + C + I$$

This model is useful when the variations in the time series are in absolute values and can be separated and traced to each of these four parts and each part can be measured independently.

5.6 MEASURES OF TRENDS

The following are the various measures of trends:

NOTES

Trend Analysis

While chance variations are difficult to identify, separate, control or predict, a more precise measurement of trend, cyclical effects and seasonal effects can be made in order to make the forecasts more reliable. In this section, we discuss techniques that would allow us to describe trend.

When a time series shows an upward or downward long-term linear trend, then regression analysis can be used to estimate this trend and project the trends into forecasting the future values of the variables involved. The equation for the straight line used to describe the linear relationship between the independent variable X and the dependent variable Y is:

$$Y = b_0 + b_1 X$$

where, b_0 = Intercept on the Y-axis and b_1 = Slope of the straight line

In time series analysis, the independent variable is time, so we will use the symbol t in place of X and we will use the symbol Y_t in place of Y_c which we have used previously.

Hence, the equation for linear trend is given as:

$$Y_t = b_0 + b_1 t$$

where, Y_t = Forecast value of the time series in period t.

b_0 = Intercept of the trend line on Y-axis.

b_1 = Slope of the trend line.

t = Time period.

As discussed earlier, we can calculate the values of b_0 and b_1 by the following formulae:

where, y = Actual value of the time series in period time t.

n = Number of periods.

$\bar{y}$ = Average value of time series .

$\bar{t}$ = Average value of .

Knowing these values, we can calculate the value of y.

Example 10.16: A car fleet owner has 5 cars which have been in the fleet for several different years. The manager wants to establish if there is a linear relationship between the age of the car and the repairs in hundreds of dollars for a given year. This way, he can predict the repair expenses for each year as the cars become older. The information for the repair costs he had collected for the last year on these cars is as follows:

Car #	*Age (t)*	*Repairs (Y)*
1	1	4
2	3	6
3	3	7
4	5	7
5	6	9

The manager wants to predict the repair expenses for the next year for the two cars that are 3 years old.

Solution: The trend in repair costs suggests a linear relationship with the age of the car, so that the linear regression equation is given as:

where

and $b_0 = \bar{y} - b_1\bar{t}$

To calculate the various values, let us form a new table as follows:

Age of Car (t)	*Repair Cost (Y)*	*tY*	t^2
1	4	4	1
3	6	18	9
3	7	21	9
5	7	35	25
6	9	54	36
Total 18	33	132	80

Y = b 5(132) (18)(33) 5 6.6 n 5(80) (18)

Knowing that $n = 5$, let us substitute these values to calculate the regression coefficients b_0 and b_1.

Then,

$$= \frac{660 - 594}{400 - 324}$$

$$= \frac{66}{76} = 0.87$$

and $b_0 = \bar{y} - b_1\bar{t}$

where

and $\bar{t} = \frac{t}{n} = \frac{18}{5} = 3.6$

NOTES

Then, b_0 6.6 0.87(3.6)

= 6.6 – 3.13

= 3.47

Hence,

The cars that are 3 years old now will be 4 years old next year, so that t = 4.

Hence,

= 6.95

Accordingly, the repair costs on each car that is 3 years old are expected to be $695.00

Smoothing Techniques

Smoothing techniques improve the forecasts of future trends provided that the time series is fairly stable with no significant trend, cyclical or seasonal effect and the objective is to smooth out the irregular component of the time series through the averaging process. There are two techniques that are generally employed for such smoothing which are as follows:

1. Moving Averages: The concept of the moving averages is based on the idea that any large irregular component of time series at any point in time will have a less significant impact on the trend, if the observation at that point in time is averaged with such values immediately before and after the observation under consideration. For example, if we are interested in computing the three-period moving average for any time period, then we will take the average of the value in such time period, the value in the period immediately preceding it and the value in the time period immediately following it. Let us illustrate this concept with the help of an example.

Example 10.17: Let the following table represent the number of cars sold in the first 6 weeks of the first two months of the year by a given dealer. Our objective is to calculate the three-week moving average.

Week	*Sales*
1	20
2	24
3	22
4	26
5	21
6	22

Solution: The moving average for the first three-week period is given as:

This moving average can then be used to forecast the sale of cars for week 4. Since the actual number of cars sold in week 4 is 26, we note that the error in the forecast is (26 – 22) = 4.

The calculation for the moving average for the next three periods is done by adding the value for week 4 and dropping the value for week 1, and taking the average for weeks 2, 3 and 4. Hence,

NOTES

Then, this is considered to be the forecast of sales for week 5. Since the actual value of the sales for week 5 is 21, we have an error in our forecast of (21 – 24) = – (3).

The next moving average for weeks 3 to 5, as a forecast for week 6 is given as:

$$\text{Moving average} = \frac{22 + 26 + 21}{3} = \frac{69}{3} = 23$$

The error between the actual and the forecast value for week 6 is (22 – 23) = – (1). (Since the actual value of the sales for week 7 is not given, there is no need to forecast such values).

Our objective is to predict the trend and forecast the value of a given variable in the future as accurately as possible so that the forecast is reasonably free from random variations. To do that, we must have the sum of individual errors, as discussed earlier, as little as possible. However, since errors are irregular and random, it is expected that some errors would be positive in value and others negative, so that the sum of these errors would be highly distorted and would be closer to zero. This difficulty can be avoided by squaring each of the individual forecast errors and then taking the average. Naturally, the minimum values of these errors would also result in the minimum value of the 'average of the sum of squared errors'. This is shown as follows:

Week	*Time Series Value*	*Moving Average*	*Error*	*Error Squared*
1	20			
2	24			
3	22			
4	26	22	4	16
5	21	24	– 3	9
6	22	23	– 1	1

Then the average of the sum of squared errors, also known as *Mean Squared Error* (MSE) is given as:

$$\text{MSE} = \frac{16 + 9 + 1}{3} = 8.67$$

The value of MSE is an often-used measure of the accuracy of the forecasting method, and the method which results in the least value of MSE is considered more accurate than others. The value of MSE can be manipulated by varying the number of data values to be included in the moving average. For example, if we had calculated the value of MSE by taking 4 periods into consideration for calculating the moving average, rather than 3, then the value of MSE would be less. Accordingly, by using trial and error method, the number of data values selected for use in forecasting would be such that the resulting MSE value would be minimum.

2. Exponential Smoothing using Least Square Method: In the moving average method, each observation in the moving average calculation receives the same weight. In other words, each value contributes equally towards the calculation of the moving average, irrespective of the number of time periods taken into consideration. In most actual situations, this is not a realistic assumption. Because of the dynamics of the

NOTES

environment over a period of time, it is more likely that the forecast for the next period would be closer to the most recent previous period than the more distant previous period, so that the more recent value should get more weight than the previous value, and so on. The exponential smoothing technique uses the moving average with appropriate weights assigned to the values taken into consideration in order to arrive at a more accurate or smoothed forecast. It takes into consideration the decreasing impact of the past time periods as we move further into the past time periods. This decreasing impact as we move down into the time period is exponentially distributed and hence, the name exponential smoothing.

In this method, the smoothed value for period *t*, which is the weighted average of that period's actual value and the smoothed average from the previous period (*t* – 1), becomes the forecast for the next period (*t* + l). Then the exponential smoothing model for time period (*t* + l) can be expressed as follows:

$$F_{(t\ \ 1)} \qquad Y_t \quad (1 \qquad)F_t$$

where $F_{(t+1)}$ = The forecast of the time series for period (*t* + 1).

Y_t = Actual value of the time series in period *t*.

α = Smoothing factor .

F_t = Forecast of the time series for period *t*.

The value of α is selected by the decision-maker on the basis of the degree of smoothing required. A small value of α means a greater degree of smoothing. A large value of α means very little smoothing. When $\alpha = 1$, then there is no smoothing at all so that the forecast for the next time period is exactly the same as the actual value of times series in the current period. This can be seen by:

when

The exponential smoothing approach is simple to use and once the value of α is selected, it requires only two pieces of information, namely Y_t and F_t to calculate .

To begin with the exponential smoothing process, we let F_t equal the actual value of the time series in period *t*, which is Y_1. Hence, the forecast for period 2 is written as:

But since we have put hence,

$$= Y_1$$

Let us now apply exponential smoothing method to the problem of forecasting car sales as discussed in the case of moving averages. The data once again is given as follows:

Week	*Time Series Value* (Y_t)
1	20
2	24
3	22
4	26
5	21
6	22

Let

Since F_2 is calculated earlier as equal to $Y_1 = 20$, we can calculate the value of F_3 as follows:

Since $F_2 = Y_1$, we get

$= 21.6$

Similar values can be calculated for subsequent periods, so that,

$$F_4 =$$
$$= 0.4(22) + 0.6(21.6)$$
$$= 8.8 + 12.96$$
$$= 21.76$$
$$F_5 = 0.4Y_4 + 0.6F_4$$
$$= 0.4(26) + 0.6(21.76)$$
$$= 10.4 + 13.056$$

$F_3 = 0.4Y_2 + 0.6F_2$ $0.4Y_3 + 0.6F_3$ $0.4(24) + 0.6(20)$ 9.6 12

$$= 23.456$$
$$F_6 = 0.4Y_5 + 0.6F_5$$
$$= 0.4(21) + 0.6(23.456)$$
$$= 8.4 + 14.07$$
$$= 22.47$$

and, $F_7 = 0.4Y_6 + 0.6F_6$

$$= 0.4(22) + 0.6(22.47)$$
$$= 8.8 + 13.48$$
$$= 22.28$$

Now we can compare the exponential smoothing forecast value with the actual values for the six time periods and calculate the forecast error.

Week	*Time Series Value* (Y_t)	*Exponential Smoothing Forecast Value* (F_t)	*Error* ($Y_t - F_t$)
1	20	–	–
2	24	20.000	4.0
3	22	21.600	0.4
4	26	21.760	4.24
5	21	23.456	– 2.456
6	22	22.470	– 0.47

(The value of F_7 is not considered because the value of Y_7 is not given).

Let us now calculate the value of MSE for this method with selected value of

From the previous table:

NOTES

Forecast errors	Squared Forecast Error
0.4	0.16
4.24	17.98
– 2.456	6.03
– 0.47	0.22
	Total = 40.39

Then,

$$\text{MSE} = 40.39/5$$
$$= 8.08$$

The previous value of MSE was 8.67. Hence, the current approach is a better one.

The choice of the value for α is very significant. Let us look at the exponential smoothing model again.

where $(Y_t - F_t)$ is the forecast error during the time period t.

The accuracy of the forecast can be improved by carefully selecting the value of α. If the time series contains substantial random variability then a small value of (known as smoothing factor or smoothing constant) is preferable. On the other hand, a larger value of would be desirable for time series with relatively little random variability $(Y_t - F_t)$.

Measuring Cyclical Effect

Cyclic variation, as we have discussed earlier, is a pattern that repeats over time periods longer than one year. These variations are generally unpredictable in relation to the time of occurrence, duration as well as amplitude. However, these variations have to be separated and identified. The measure we use to identify cyclical variation is the *percentage of trend* and the procedure used is known as the *residual trend.*

As we have discussed earlier, there are four components of time series. These are secular trend (T), seasonal variation (S), cyclical variation (C) and irregular (or chance) variation (I). Since the time period considered for seasonal variation is less than one year, it can be excluded from the study, because when we look at time series consisting of annual data spread over many years, then only the secular trend, cyclical variation and irregular variation are considered.

Since secular trend component can be described by the trend line (usually calculated by line of regression), we can isolate cyclical and irregular components from the trend. Furthermore, since irregular variation occurs by chance and cannot be predicted or identified accurately, it can be reasonably assumed that most of the variation in time series left unexplained by the trend component can be explained by

NOTES

the cyclical component. In that respect, cyclical variation can be considered as the *residual*, once other causes of variation have been identified.

The measure of cyclic variation as percentage of trend is calculated as follows:

(1) Determine the trend line (usually by regression analysis).

(2) Compute the trend value Y_t for each time period (t) under consideration.

(3) Calculate the ratio Y/Y_t for each time period.

(4) Multiply this ratio by 100 to get the percentage of trend, so that:

$$\text{Percentage of trend} = \left(\frac{Y}{Y_t}\right)100$$

Free Hand Curve Method

This is a simple method of studying trends. In this method the given time series data are plotted on graph paper by taking time on X-axis and the other variable on Y-axis. The graph obtained will be irregular as it would include short-run oscillations. We may observe the up and down movement of the curve and if a smooth freehand curve is drawn passing approximately all points of a curve previously drawn, it would eliminate the short-run oscillations (seasonal, cyclical and irregular variations) and show the long-period general tendency of the data.

This is exactly what is meant by t**rend**. However, it is very difficult to draw a freehand smooth curve and different persons are likely to draw different curves from the same data. The following points must be kept in mind in drawing a freehand smooth curve:

1. That the curve is smooth.
2. That the numbers of points above the line or curve are equal to the points below it.
3. That the sum of vertical deviations of the points above the smoothed line is equal to the sum of the vertical deviations of the points below the line. In this way the positive deviations will cancel the negative deviations. These deviations are the effects of seasonal cyclical and irregular variations and by this process they are eliminated.
4. The sum of the squares of the vertical deviations from the trend line curve is minimum. This is one of the characteristics of the trend line fitted by the method of least squares.

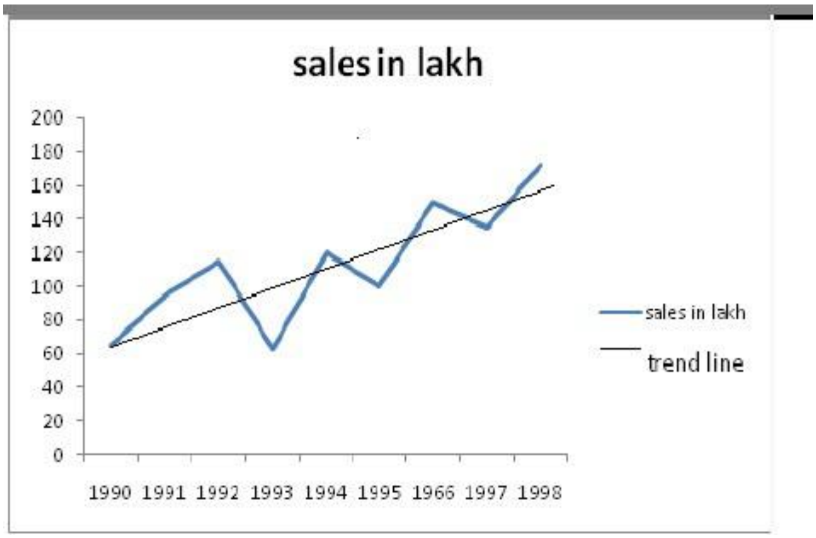
The trend values can be read for various time periods by locating them on the trend line against each time period. The following example will illustrate the fitting of a freehand curve to set of time series values:

For example: The table below shows the data of sale of nine years:

Year	1990	1991	1992	1993	1994	1995	1996	1997	1998
Sales in (lakh units)	65	95	115	63	120	100	150	135	172

If we draw a graph taking year on x-axis and sales on y- axis, it will be irregular as shown below. Now drawing a freehand curve passing approximately through all this points will represent trend line (shown below by black line).

NOTES



Merits

The following are the merits of free hand curve method:

1. It is simple method of estimating trend which requires no mathematical calculations.
2. It is a flexible method as compared to rigid mathematical trends and, therefore, a better representative of the trend of the data.
3. This method can be used even if trend is not linear.
4. If the observations are relatively stable, the trend can easily be approximated by this method.
5. Being a non mathematical method, it can be applied even by a common man.

Demerits

The following are the demerits of free hand curve method:

1. It is subjective method. The values of trend, obtained by different statisticians would be different and hence, not reliable.
2. Predictions made on the basis of this method are of little value.

5.7 SCOPE IN BUSINESS

Seasonal variation has been defined as predictable and repetitive movement around the trend line in a period of one year or less. For the measurement of seasonal variation, the time interval involved may be in terms of days, weeks, months or quarters. Because of the predictability of seasonal trends, we can plan in advance to meet these variations. For example, studying the seasonal variations in the production data makes it possible to plan for hiring of additional personnel for peak periods of production or to accumulate an inventory of raw materials or to allocate vacation time to personnel, and so on.

In order to isolate and identify seasonal variations, we first eliminate, as far as possible the effects of trend, cyclical variations and irregular fluctuations on the time series. Some of the methods used for the measurement of seasonal variations are described as follows.

Check Your Progress

9. What do you mean by the term 'trend'?
10. What is the meaning of cyclical fluctuation?
11. What factors cause seasonal variations?

NOTES

Simple Averages

This is the simplest method of isolating seasonal fluctuations in time series. It is based on the assumption that the series contain only the seasonal and irregular fluctuations. Assume that the time series involve monthly data over a time period of, say, 5 years. Assume further that we want to find the seasonal index for the month of March. (The seasonal variation will be the same for March in every year. Seasonal index describes the degree of seasonal variation).

Then, the seasonal index for the month of March will be calculated as follows:

$$\text{Seasonal Index for March} \quad \frac{\text{Monthly average for March}}{\text{Average of monthly averages}} \quad 10$$

The following steps can be used in the calculation of seasonal index (variation) for the month of March (or any month), over the five years period, regarding the sale of cars by one distributor.

1. Calculate the average sale of cars for the month of March over the last 5 years.
2. Calculate the average sale of cars for each month over the 5 years and then calculate the average of these monthly averages.
3. Use the formula to calculate seasonal index for March.

Let us say that the average sale of cars for the month of March over the period of 5 years is 360, and the average of all monthly average is 316. Then the seasonal index for March = (360/316) × 100 = 113.92.

Moving Averages

This is the most widely used method of measuring seasonal variations. The seasonal index is based upon a mean of 100 with the degree of seasonal variation (seasonal index) measured by variations away from this base value. For example, if we look at the seasonality of rental of row boats at the lake during the three summer months (a quarter) and we find that the seasonal index is 135 and we also know that the total boat rentals for the entire last year was 1680, then we can estimate the number of summer rentals for the row boats.

The average number of quarterly boats rented = 1680/4 = 420.

The seasonal index, 135 for the summer quarter means that the summer rentals are 135 percent of the average quarterly rentals.

Hence, summer rentals = 420 × (135/100) = 567.

The steps required to compute the seasonal index can be enumerated by illustrating an example.

Example 10.18: Assume that a record of rental of row boats for the previous 3 years on a quarterly basis is given as follows:

Year	*Rentals per quarter*				*Total*
	I	II	III	IV	
1991	350	300	450	400	1500
1992	330	360	500	410	1600
1993	370	350	520	440	1680

NOTES

Solution:

Step 1. The first step is to calculate the four-quarter moving total for time series. This total is associated with the middle data point in the set of values for the four quarters, shown as follows.

Year	*Quarters*	*Rentals*	*Moving Total*
1991	I	350	
	II	300	
			1500
	III	450	
	IV	400	

The moving total for the given values of four quarters is 1500 which is simply the addition of the four quarter values. This value of 1500 is placed in the middle of values 300 and 450 and recorded in the next column. For the next moving total of the four quarters, we will drop the value of the first quarter, which is 350, from the total and add the value of the fifth quarter (in other words, first quarter of the next year), and this total will be placed in the middle of the next two values, which are 450 and 400, and so on. These values of the moving totals are shown in column 4 of the following table.

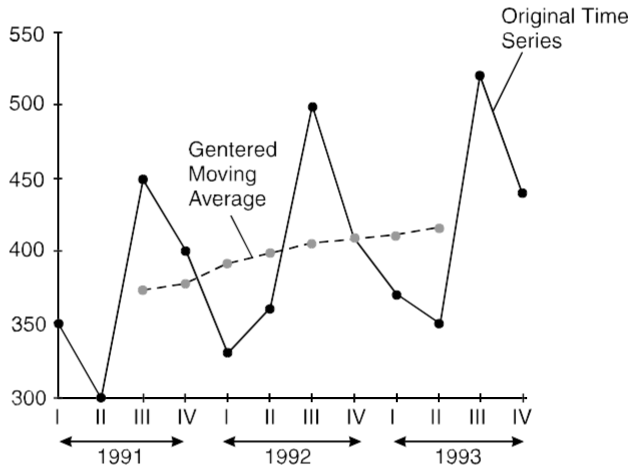
Step 2. The next step is to calculate the quarter moving average. This can be done by dividing the four quarter moving total, as calculated in Step 1 earlier, by 4, since there are 4 quarters. The quarter moving average is recorded in column 5 in the table. The entire table of calculations is shown as follows:

Year	*Quarters*	*Rentals*	*Quarter Moving Total*	*Quarter Moving Average*	*Quarter Centered Moving Average*	*Percentage of Actual to Centered Moving Average*
(1)	(2)	(3)	(4)	(5)	(6)	(7)
	I	350				
	II	300				
			1500	375.0		
	III	450			372.50	120.80
			1480	370.0		
	IV	400			377.50	105.96
			1540	385.0		
1992	I	330			391.25	84.35
			1590	397.5		
	II	360			398.75	90.28
			1600	400.0		
	III	500			405.00	123.45
			1640	410.0		
	IV	410			408,75	100.30
			1630	407.5		
1993	I	370			410.00	90.24
			1650	412.5		
	II	350			416.25	84.08
			1680	420.0		
	III	520				
	IV	440				

Step 3. After the moving averages for each of the consecutive four quarters have been taken, we centre these moving averages. As we see from the table, the quarterly moving average falls between the quarters. This is because the number of quarters is even which is 4. If we had odd number of time periods, such as 7 days of the week, then the moving average would already be centered and the third step here would not be necessary. Accordingly, we centre our averages in order to associate each average with the corresponding quarter, rather than between the quarters. This is shown in column 6, where the centered moving average is calculated as the average of the two consecutive moving averages.

The moving average (or the centered moving average) aims to eliminate seasonal and irregular fluctuations (S and I) from the original time series, so that this average represents the cyclical and trend components of the series.

As the following graph shows the centered moving average has smoothed the peaks and troughs of the original time series.



Step 4. Column 7 in the table contains calculated entries which are percentages of the actual values to the corresponding centered moving average values. For example, the first four quarters centered moving average of 372.50 in the table has the corresponding actual value of 450, so that the percentage of actual value to centered moving average would be:

$$\frac{\text{Actual Value}}{\text{Centered Moving Average Value}} \times 100$$

$$= \frac{450}{372.5} \times 100$$

$$= 120.80$$

Step 5. The purpose of this step is to eliminate the remaining cyclical and irregular fluctuations still present in the values in Column 7 of the table. This can be done by calculating the 'modified mean' for each quarter. The modified mean for each quarter of the three years time period under consideration, is calculated as follows.

(a) Make a table of values in column 7 of the previous table (percentage of actual to moving average values) for each quarter of the three years as shown in the following table.

NOTES

Year	Quarter I	Quarter II	Quarter (III)	Quarter (IV)
1991	–	–	120.80	105.96
1992	84.35	90.28	123.45	100.30
1993	90.24	84.08	–	–

(b) We take the average of these values for each quarter. It should be noted that if there are many years and quarters taken into consideration instead of 3 years as we have taken, then the highest and lowest values from each quarterly data would be discarded and the average of the remaining data would be considered. By discarding the highest and lowest values from each quarter data, we tend to reduce the extreme cyclical and irregular fluctuations, which are further smoothed when we average the remaining values. Thus, the modified mean can be considered as an index of seasonal component. This modified mean for each quarter data is shown as follows:

$$\text{Quarter I} = \frac{84.35 + 90.24}{2} = 87.295$$

$$\text{Quarter II} = \frac{90.28 + 84.08}{2} = 87.180$$

$$\text{Quarter III} = \frac{120.80 + 123.45}{2} = 122.125$$

Total = 399.73

The modified means as calculated here are preliminary seasonal indices. These should average 100 percent or a total of 400 for the 4 quarters. However, our total is 399.73. This can be corrected by the following step.

Step 6. First, we calculate an adjustment factor. This is done by dividing the desired or the expected total of 400 by the actual total obtained of 399.73, so that:

$$\text{Adjustment} = \frac{400}{399.73} = 1.0007$$

By multiplying the modified mean for each quarter by the adjustment factor, we get the seasonal index for each quarter, so that:

Quarter I = 87.295 × 1.0007 = 87.356

Quarter II = 87.180 × 1.0007 = 87.241

Quarter III = 122.125 × 1.0007 = 122.201

Quarter IV = 103.13 × 1.0007 = 103.202

Total = 400.000

$$\text{Average seasonal index } = \frac{400}{4} = 100$$

(This average seasonal index is approximated to 100 because of rounding-off errors).

The logical meaning behind this method is based on the fact that the centered moving average part of this process eliminates the influence of secular trend and cyclical fluctuations ($T \times C$). This may be represented by the following expression:

NOTES

$$\frac{T \times S \times C \times I}{T \times C} = S \times I$$

where ($T \times S \times C \times I$) is the influence of trend, seasonal variations, cyclic fluctuations and irregular or chance variations.

Thus, the ratio to moving average represents the influence of seasonal and irregular components. However, if these ratios for each quarter over a period of years are averaged, then the most random or irregular fluctuations would be eliminated so that,

$$\frac{S \times I}{I} = S$$

and this would give us the value of seasonal influences.

Measuring Irregular Variation and Seasonal Adjustments

Typically, irregular variation is random in nature, unpredictable and occurs over comparatively short periods of time. Because of its unpredictability, it is generally not measured or explained mathematically. Usually, subjective and logical reasoning explains such variation that occurs; for example, long period of cold weather in Brazil and Columbia results in increase in the price of coffee beans, because cold weather destroys coffee plants. Similarly, the Persian Gulf War, an irregular factor resulted in increase in airline and ship travel for a number of months because of the movement of personnel and supplies. However, the irregular component can be isolated by eliminating other components from the time series data. For example, time series data contains ($T \times S \times C \times I$) components and if we can eliminate ($T \times S \times C$) elements from the data, then we are left with (I) component. We can follow the previous example to determine the (I) component as follows. The data presented has already been earlier provided or calculated.

Year	*Quarters*	*Rentals Time Series Values ($T \times S \times C \times I$)*	*Centered Moving Average ($T \times C$)*	*$T \times S \times C \times I / (T \times C) = S \times I$*
1991	I	350	–	–
	II	300	–	–
	III	450	372.50	1.208
	IV	400	377.50	1.060
1992	I	330	391.25	0.843
	II	360	398.75	0.903
	III	500	405.00	1.235
	IV	410	408.75	1.003
1993	I	370	410.00	0.902
	II	350	416.25	0.841
	III	520	–	–
	IV	440	–	–

The seasonal indices for each quarter have already been calculated as:

Quarter I = 87.356

Quarter II = 87.241

NOTES

Quarter III = 122.201

Quarter IV = 103.202

Then the seasonal influence is given by:

Quarter I = 87.356/100 = .874

Quarter II = 87.241/100 = .872

Quarter III = 122.201/100 = 1.222

Quarter IV = 103.202/100 = 1.032

Making another table of ($S \times I$) values and (S) values and dividing ($S \times I$) by (S) we get the values of (I) as follows:

Year	*Quarters*	*(S × I)*	*(S)*	*(I)*
1991	I	–	–	–
	II	–	–	–
	III	1.208	1.222	0.988
	IV	1.060	1.032	1.027
1992	I	0.843	0.874	0.965
	II	0.903	0.872	1.036
	III	1.235	1.222	1.011
	IV	1.003	1.032	0.972
1993	I	0.902	0.874	1.032
	II	0.841	0.872	0.964
	III	–	–	–
	IV	–	–	–

Seasonal Adjustments

Many times we read about time series values as seasonally adjusted. This is accomplished by dividing the original time series values by their corresponding seasonal indices. These deseasonalized values allow more direct and equitable comparisons of values from different time periods. For example, in comparing the demands for rental row boats (example that we have been following), it would not be equitable to compare the demand of second quarter (spring) with the demand of third quarter (summer), when the demand is traditionally higher. However, these demand values can be compared when we remove the seasonal influence from these time series values.

The seasonally adjusted values for the demand of row boats in each quarter are based on the values previously calculated and is shown as follows.

Year	*Quarter*	*Rentals (T × S × C × I)*	*(S)*	*Seasonally Adjusted Values*	*Rounded-off Values*
1991	I	350	–	–	–
	II	300	–	–	–
	III	450	1.222	368.25	368
	IV	400	1.032	387.60	388

Contd...

1992	I	330	0.874	377.57	378
	II	360	0.872	412.80	413
	III	500	1.222	409.16	409
	IV	410	1.032	397.29	397
1993	I	370	0.874	423.34	423
	II	350	0.872	401.38	401
	III	520	–	–	–
	IV	440	–	–	–

The seasonally adjusted value for each quarter is calculated as:

$$\frac{\text{Original Value}}{\text{Seasonal Index}}$$

These calculations complete the process of separating and identifying the four components of the time series, namely secular trend (*T*), seasonal variation (*S*), cyclical variation (*C*) and irregular variation (*I*).

5.8 SUMMARY

5.8.1 Index numbers are a specialized type of average. They are designed to measure the relative change in the level of a phenomenon with respect to time, geographi-cal locations or some other characteristics.

5.8.2 According to Wheldon, 'Index number is a statistical device for indicating the relative movements of the data where measurement of actual movements is difficult or incapable of being made.'

- According to F.Y. Edgeworth, 'Index number shows by its variations the changes in a magnitude which is not susceptible either of accurate measurement in itself or of direct valuation in practice.'
- Originally, the index numbers were developed for measuring the effect of changes in the price level. But today the index numbers are also used to measure changes in industrial production, fluctuations in the level of business activities or variations in the agricultural output, etc.
- In the words of G. Simpson and F. Kafka: 'Index numbers are today one of the most widely used statistical devices. They are used to take the pulse of the economy and they have come to be used as indicators of inflationary or deflationary tendencies.'
- Methods of constructing index numbers can broadly be divided into two classes namely, unweighted indices and weighted indices.
- In case of unweighted indices, weights are not expressly assigned, whereas in the weighted indices weights are expressly assigned to the various items. Each of these types may be further classified under two heads as aggregate of prices method and average of price relative's method.
- Weighted aggregate of prices index is similar to the simple aggregative type with the fundamental difference that weights are assigned explicitly to the various

NOTES

Check Your Progress

12. Define the term irregular or random variation.
13. What are the two methods adopted in smoothing techniques?
14. List the methods used to measure seasonal variation.

items included in the index.

NOTES

- In Laspeyre's method, base year quantities are taken as weights. The formula for constructing the index is, .
- Laspeyre's index is very widely used. It tells us about the change in the aggregate value of the base period list of goods when valued at a given period price.
- In Paasche's index method, the current year quantities (q_1) are taken as weights. The formula for constructing this index is, $P_{01} = \frac{\sum P_1 q_1}{\sum P_0 q_1} \times 100$.
- The Bowley-Drobisch method is the simple arithmetic mean of Laspeyre's and Paasche's indices. The formula for constructing Bowley-Drobisch index is, $P_{01} = \frac{L + P}{2}$.
- It is absolutely necessary that the purpose of the index numbers be rigourously defined. This would help in deciding the nature of data to be collected, the choice of the base year, the formula to be used and other related matters.
- The base year should not be too distant in the past. Since the index numbers are useful in decision-making, and economic practices are often a matter of the short run, we should choose a base which is relatively close to the year being studied.
- While selecting the base year, a decision has to be made whether the base shall remain fixed or not. If the period of comparison is fixed for all current years, it is called fixed base method. If, on the other hand, the prices of the current year are linked with the prices of the preceding year and not with the fixed year or period, it is called chain base method.
- Chain base method is useful in cases where there are quick and frequent changes in fashion, tastes and habits of the people. In such cases comparison with the preceding year is more worthwhile.
- According to Prof. Fisher, the formula for constructing the index number should permit not only the interchange of the two times without giving inconsistent results, it should also permit the interchange of weights without giving inconsistent results.
- A cost of living index is a theoretical price index that measures relative cost of living over time or regions. It is an index that measures differences in the price of goods and services, and allows for substitutions to other items as prices vary.
- Accurate forecasting is an essential element of planning of any organization or policy. This requires studying previous performances in order to forecast future activities.
- When a projection of the pattern of future economic activity is known and the level of future business activity is understood, the desirability of an alternative course of action and the selection of an optimum alternative can be examined and forecast.
- The quality of such forecasts is strongly related to the relevant information that can be extracted from past data.

- Time series analysis method helps in making accurate predictions and also in situations where the future is expected to be similar to or at least predictive from the past.

5.9 KEY TERMS

5.9.1 **Index numbers:** It measures the relative change in the magnitude of a group of related, distinct variables in two or more situations. Index numbers can be used to measure changes in prices, wages production, employment, national income, etc., over a period of time

5.9.2 **Value index number:** It compares the total value of all commodities in the current period with the total value in the base period

5.9.3 **Volume index numbers:** These numbers measure the changes in physical volumes of goods produced, distributed or consumed

5.9.4 **Base period:** It refers to the point of reference established in the construction ofindex numbers of prices

5.9.5 **Splicing:** It is the procedure employed for connecting an old series of index numbers with a revised series in order to make the series continuous

5.9.6 **Deflation:** It is defined as the sustained fall in the general price level

5.9.7 **Seasonal variation:** It involves patterns of change that repeat over a period of one year or less. The factors that cause seasonal variations are season, climate, customs and festivals

5.9.8 **Irregular variation:** These variations are unpredictable and can be accidental,random or simply due to chance

5.9.9 **Cyclic variation:** It is a pattern that repeats over time periods longer than oneyear

5.10 ANSWERS TO 'CHECK YOUR PROGRESS'

1. The two types of methods for constructing index numbers are:
   (a) Unweighted indices
   (b) Weighted indices
2. The methods for constructing weighted aggregate of price index are (i) Laspeyres' method (ii) Paasche's method (iii) Bowley-Drobish method (iv) Marshall-Edgeworth method (v) Fisher's ideal index (vi) Kellys' method
3. While selecting the base year to determine index, a decision has to be made whether the base shall be fixed or not. If the period of comparison is not fixed for all current years and the prices of the current year are linked with the prices of the preceding year, it is called chain base method.
4. The term 'weight' refers to the relative importance of different items in the

construction of the index.

5. A value index V is the sum of the value of a given year divided by sum of the values of the base year. In simple terms, the formula for value can thus be stated as $V = \frac{V_1}{V_0}$. Here, both price and quantity are variable in the numerator.

6. The procedure employed for connecting an old series of index numbers with a revised series in order to make the series continuous is called splicing.
7. The process of calculating the real wages by applying index numbers to the money wages so as to allow for the change in the price level is called deflating. It is a way by which a series of money wages or incomes can be corrected for price changes to find out the level of real wages or income.
8. The uses of index numbers are (i) They help in framing suitable policies (ii) They help in studying trends and tendencies (iii) They are useful in deflating.
9. The term trend means the general long-term movement in the time series value of the variable (*Y*) over a fairly long period of time. Here, '*Y*' stands for such factors like sales, population and crime rate that we are interested in evaluating for the future.
10. Regular swings or patterns that repeat over a long period of time are known as cyclical fluctuations. These are usually unpredictable in relation to the time of occurrence, the duration as well as the amplitude.
11. Factors like changes in climate and weather, and customs and traditions cause seasonal variations.
12. Those variations which are accidental, random or occur due to chance factors, are known as irregular or random variations.
13. The two methods adopted in smoothing techniques are:
    - Moving Averages
    - Exponential Smoothing
14. The methods used in measuring seasonal variation are:
    - Simple Average Method
    - Moving Average Method

5.11 QUESTIONS AND EXERCISES

Short-Answer Questions

1. What is the importance of using index numbers?
2. What are the various uses of index numbers?
3. How is index number constructed?
4. What are fixed base and chain base methods.
5. Name the methods used in assigning weights.
6. How will you obtain price quotations?
7. What is Paasche's index? How is it calculated?
8. Define Fisher's ideal index. Write its formula and state why it is called ideal?
9. What are the steps involved in the construction of weighted average of price relatives index?
10. Write the different formulas for calculating quantity index.
11. Differentiate between secular trend and cyclic fluctuation.
12. How is irregular variation caused?

13. Define the term seasonal variation.
14. What do you mean by trend analysis?
15. How will you measure cyclical effect?
16. What are the ways to measure irregular variation?
17. How are seasonal adjustments made?

Long-Answer Questions

1. (a) 'An index number is a special type of average.' Discuss.

   (b) What points should be taken into consideration in the construction of index numbers? Discuss.
2. What is meant by 'weighting' in statistics? Why is it necessary to assign weights in the construction of index numbers? What are the various ways of assigning weights in index number construction?
3. Explain briefly time reversal and factor reversal tests of index numbers. Indicate whether the following index numbers satisfy one or the other of these tests: Laspeyre's, Paasche's, Marshall-Edgeworth's and Fisher's Ideal Index Numbers.
4. What is meant by deflating? What purpose does it serve? Explain briefly the procedure for deflating with the help of an example.
5. What is base shifting? Why does it become necessary to shift the base of index numbers? Give an example of the shifting of base of index numbers.
6. From the following data, compute the index number of prices for the year 1980 with 1979 as base, using:

   (a) Laspeyre's method (b) Paasche's method

   (c) Bowley-Drobisch method (d) Fisher's ideal formula

	1979		1980	
Commodities	*Price*	*Quantity*	*Price (`)*	*Quantity*
A	20	8	40	6
B	50	10	60	5
C	40	15	50	10
D	20	20	20	15

7. An enquiry into the budgets of middle class families of a certain city revealed that on an average the percentage expenses on the different groups were—Food—45, Rent—15, Clothing—12, Fuel and Light—8 and Miscellaneous—20. The group index numbers for the current year as compared with a fixed base period were respectively 410, 150, 343, 248 and 285. Calculate the consumer price index number for the current year. Mr. X was getting ` 240 p.m. in the base period and ` 430 p.m. in the current year. Calculate his real income in the current year. State how much he ought to have received as extra allowance to maintain his former standard of living.
8. From the following chain base index numbers, prepare fixed base index numbers with base 1970 = 100.

Year	1971	1972	1973	1974	1975
Index	110	160	140	200	150

NOTES

9. A price index series was started with 1961 as base. By 1965, it rose by 20 per cent, the link relative for 1966 was 90. In this year, a new series was started. This new series rose by 12 points by next year. But during the next three years the rise was not rapid. During 1970, the price level was only 10 per cent higher than that of 1967. Splice the two series and calculate the index number for various years by shifting the base to 1967.

10. The following data shows the number of Lincoln Continental cars sold by a dealer in Queens during the 12 months of 1994.

Month	*Number Sold*
Jan	52
Feb	48
Mar	57
Apr	60
May	55
June	62
July	54
Aug	65
Sept	70
Oct	80
Nov	90
Dec	75

(a) Calculate the three-month moving average for this data.

(b) Calculate the five-month moving average for this data.

(c) Which one of these two moving averages is a better smoothing technique and why?

11. The owner of six gasoline stations in New Jersey would like to have some reasonable indication of future sales. He would like to use the moving average method to forecast future sales. He has recorded the quarterly gasoline sales (in thousands of gallons) for all his gas stations for the past three years. These are shown in the following table.

Year	*Quarter*	*Sales*
1	1	38
	2	58
	3	80
	4	30
2	1	40
	2	60
	3	50
	4	55
3	1	50
	2	45
	3	80
	4	70

(a) Calculate the three-quarter moving average.

(b) Calculate the five-quarter moving average.

(c) Plot the quarterly sales and also both the moving averages on the same graph. Which of these two moving average seems to be a better smoothing technique?

NOTES

12. An economist has calculated the variable rate of return on money market funds for the last twelve months as follows:

Month	*Rate of Return (%)*
January	6.2
February	5.8
March	6.5
April	6.4
May	5.9
June	5.9
July	6.0
August	6.8
September	6.5
October	6.1
November	6.0
December	6.0

(a) Using a three-month moving average, forecast the rate of return for next January.

(b) Using exponential smoothing method and setting, α = 0.8, forecast the rate of return for next January.

13. The Indian Motorcycle Company is concerned about declining sales in the western region. The following data shows monthly sales (in millions of dollars) of the motorcycles for the past twelve months.

Month	*Sales*
January	6.5
February	6.0
March	6.3
April	5.1
May	5.6
June	4.8
July	4.0
August	3.6
September	3.5
October	3.1
November	3.0
December	3.0

NOTES

(a) Plot the trend line and describe the relationship between sales and time.

(b) What is the average monthly change in sales?

(c) If the monthly sales fall below $2.4 million, then the West Coast office must be closed. Is it likely that the office will be closed during the next six months?

14. An institution dealing with pension funds is interested in buying a large block of stock of Azumi Business Enterprises (ABE). The president of the institution has noted down the dividends paid out on common stock shares for the last ten years. This data is presented as follows:

Year	*Dividend ($)*
1985	3.20
1986	3.00
1987	2.80
1988	3.00
1989	2.50
1990	2.10
1991	1.60
1992	2.00
1993	1.10
1994	1.00

(a) Plot the data.

(b) Determine the value of regression coefficients.

(c) Estimate the dividend expected in 1995.

(d) Calculate the points on the trend line for the years 1987 and 1991 and plot the trend line.

15. Rinkoo Camera Corporation has ten camera stores scattered in five areas of New York city. The president of the company wants to find out if there is any connection between the sales price and the sales volume of Nikon F-1 camera in the various retail stores. He assigns different prices of the same camera for the different stores and collects data for a thirty-day period. The data is presented as follows. The sales volume is in number of units and the price is in dollars.

Store	*Price*	*Volume*
1	550	420
2	600	400
3	625	300
4	575	400
5	600	340
6	500	440
7	450	500
8	480	460
9	550	400
10	650	310

(a) Plot the data.

(b) Estimate the linear regression of sales on price.

(c) What effect would you expect on sales if the price of the camera in store number 7 is increased to $530?

(d) Calculate the points on the trend line for stores 4 and 7 and plot the trend line.

16. The following data shows the sales revenues for sales of used cars sold by Atlantic Company for the months of January to April in 1995.

Month	*Sales ($'00,000)*
January	95
February	105
March	100
April	110

Find the error between the actual value and the forecast value for the months of February, March and April of 1995, using exponential smoothing method with = 0.6.

17. The following data presents the rate of unemployment in South India for 12 years from 1982 to 1993.

Year	*Per cent Unemployed*
1982	12.6
1983	12.2
1984	13.0
1985	13.5
1986	12.8
1987	12.7
1988	13.1
1989	13.6
1990	13.5
1991	13.8
1992	14.2
1993	14.0

α

(a) Smooth out the fluctuations using a four-year moving average.

(b) Use the exponential smoothing model to forecast the unemployment rate in South India for the year 1995. Assume $\alpha = 0.4$.

(c) Calculate the value of MSE.

15. Juhu Chawla has a car dealership for Toyota in Bombay along with her sister Ammu. The number of cars sold for the first 7 months of 1995 are as follows:

Month	*Cars Sold*
Jan	45
Feb	52
Mar	41
Apr	36
May	49
June	47
July	43

NOTES

Juhu wants to predict the car sales for the month of August by using exponential smoothing method with an value of 0.4. Her sister thinks that an value of 0.8 would be more suitable.

What is the forecast in each case and who do you think is more correct based on these given values?

19. A restaurant manager has recorded the daily number of customers for four weeks. He wants to improve customer service and change employee scheduling as necessary based on the expected number of daily customers in the future. The following data represent the daily number of customers as recorded by the manager for the four weeks.

Week	*Mon*	*Tues*	*Wed*	*Thurs*	*Fri*	*Sat*	*Sun*
1	440	400	480	510	650	800	710
2	510	430	500	520	740	850	800
3	490	480	410	630	720	810	690
4	500	500	470	540	780	900	850

Determine the daily seasonal indices using the seven-day moving average.

20. The Department of Health has compiled data on the liquor sales in the United States (in billion dollars) for each quarter of the last four years. This quarterly data is given in the following table.

Year	*Quarter*	*Sales*
1991	I	4.5
	II	4.8
	III	5.0
	IV	6.0
1992	I	4.0
	II	4.4
	III	4.9
	IV	5.8
1993	I	4.2
	II	4.6
	III	5.2
	IV	6.1
1994	I	4.5
	II	4.6
	III	4.9
	IV	5.5

(a) Using moving average method, find the values of combined trend and cyclical component.

(b) Find the values of combined seasonal and irregular component.

(c) Find the values of the seasonal indices for each quarter.

(d) Find the seasonally adjusted values for the time series.

(e) Find the value of the irregular component.

NOTES

21. A real estate agency has been in business for the last 4 years and specializes in the sales of 2-family houses. The sales in the last 4 years have grown from 20 houses in the first year to 105 houses last year. The owner of the agency would like to develop a forecast for sale of houses in the coming year. The quarterly sales data for the last 4 years are shown as follows.

Year	*Quarter(1)*	*Quarter(2)*	*Quarter(3)*	*Quarter(4)*
1	8	6	2	4
2	10	8	8	12
3	18S	12	15	25
4	25	20	28	32

(a) Using moving average method, find the values of combined trend and cyclical component.

(b) Find the values of combined seasonal and irregular component.

(c) Compute the seasonal indices for the four quarters.

(d) Deseasonalize the data and use the deseasonalized time series to identify the trend.

(e) Find the value of irregular component.

5.12 FURTHER READING

Allen, R.G.D. 2008. *Mathematical Analysis For Economists*. London: Macmillan and Co., Limited.

Chiang, Alpha C. and Kevin Wainwright. 2005. *Fundamental Methods of Mathematical Economics*, 4 edition. New York: McGraw-Hill Higher Education.

Yamane, Taro. 2012. *Mathematics For Economists: An Elementary Survey*. USA: Literary Licensing.

Baumol, William J. 1977. *Economic Theory and Operations Analysis*, 4th revised edition. New Jersey: Prentice Hall.

Hadley, G. 1961. *Linear Algebra*, 1st edition. Boston: Addison Wesley.

Vatsa , B.S. and Suchi Vatsa. 2012. *Theory of Matrices*, 3rd edition. London: New Academic Science Ltd.

Madnani, B C and G M Mehta. 2007. *Mathematics for Economists*. New Delhi: Sultan Chand & Sons.

Henderson, R E and J M Quandt. 1958. *Microeconomic Theory: A Mathematical Approach*. New York: McGraw-Hill.

Nagar, A.L. and R.K.Das. 1997. *Basic Statistics*, 2nd edition. United Kingdom: Oxford University Press.

Gupta, S.C. 2014. *Fundamentals of Mathematical Statistics*. New Delhi: Sultan Chand & Sons.

M. K. Gupta, A. M. Gun and B. Dasgupta. 2008. *Fundamentals of Statistics*. West Bengal: World Press Pvt. Ltd.

NOTES

Saxena, H.C. and J. N. Kapur. 1960. *Mathematical Statistics*, 1st edition. New Delhi: S. Chand Publishing.

Hogg, Robert V., Joeseph McKean and Allen T Craig. *Introduction to Mathematical Statistics*, 7th edition. New Jersey: Pearson.

Institute of Distance Education

Rajiv Gandhi University

A Central University

Rono Hills, Arunachal Pradesh

Contact us:

+91-98638 68890 Ide Rgu Ide Rgu helpdesk.ide@rgu.ac.in

Institute of Distance Education
Rajiv Gandhi University

A Central University

Rono Hills, Arunachal Pradesh

Contact us:

+91-98638 68890

Ide Rgu

Ide Rgu

helpdesk.ide@rgu.ac.in