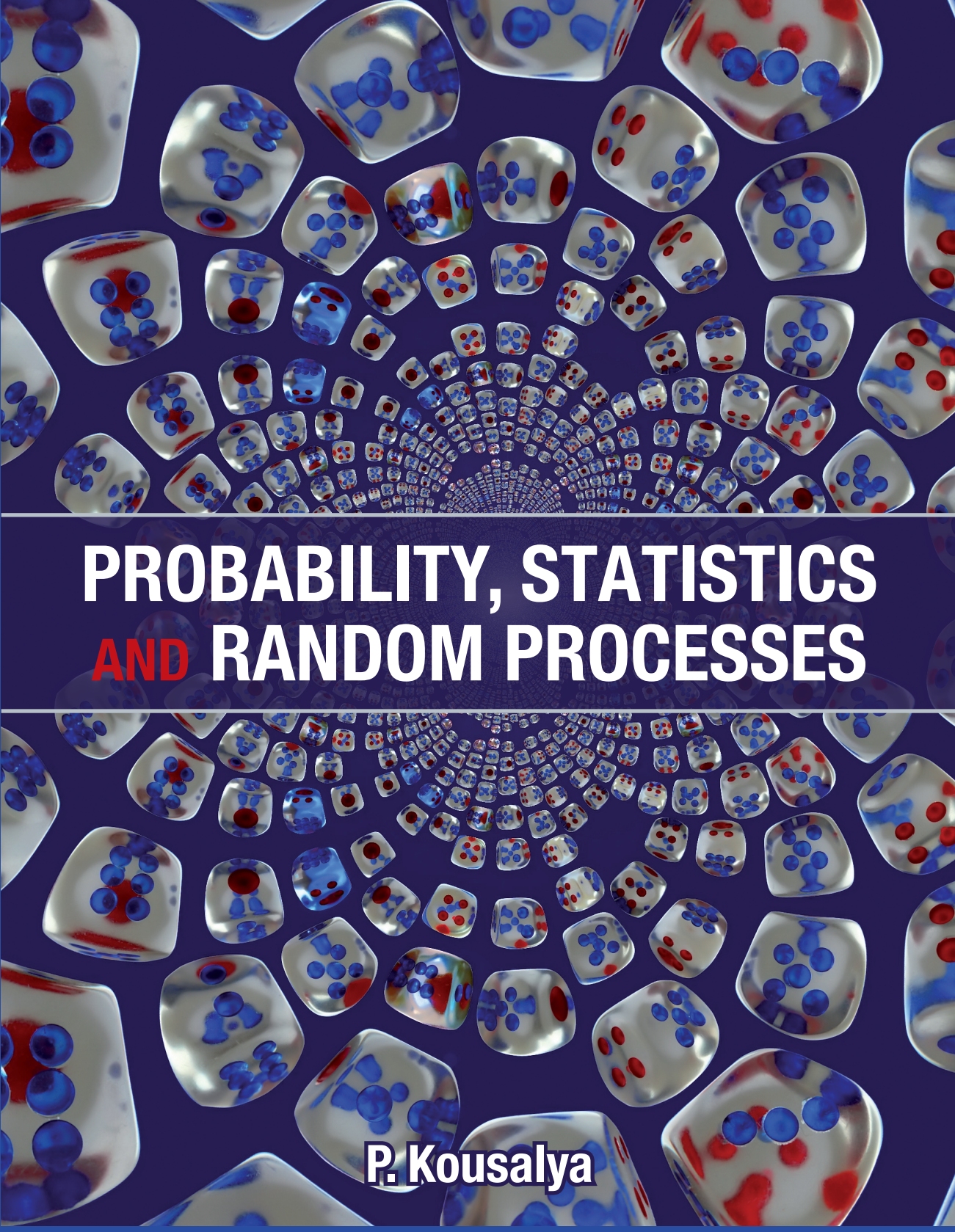


PROBABILITY, STATISTICS AND RANDOM PROCESSES

P. Kousalya

ALWAYS LEARNING

PEARSON

Probability, Statistics and Random Processes

This page is intentionally left blank.

Probability, Statistics and Random Processes

P. KOUSALYA

Professor
Department of Humanities and Sciences (Mathematics)
Vignana Bharathi Institute of Technology
Ghatkesar, Hyderabad

PEARSON

Chennai • Delhi

ISBN 9788131774526
eISBN 9789332514140

Head Office: A-8(A), Sector 62, Knowledge Boulevard, 7th Floor, NOIDA 201 309, India
Registered Office: 11 Local Shopping Centre, Panchsheel Park, New Delhi 110 017, India

Contents

Preface

Probability, Statistics and Random Process is an important book for not only engineering students but any postgraduate student of sciences. For several years, I had been teaching a course on calculus-based probability and statistics mainly for mathematics, science, and engineering students. While there were plenty of books covering one area or the other, it was surprisingly difficult to find one that covered the subject both in a satisfying way and on the appropriate level of difficulty. As I changed text books often, plenty of lecture notes got accumulated and it seemed like a good idea to organize them into a textbook.

The book is useful not only to engineering students (ECE, Mechanical, Civil, CSE, IT, and many other branches of engineering) but also to MCA and MSc (Mathematics and Statistics) students and to many other undergraduate and postgraduate students of various universities in India. The book is written in such a manner that beginners may develop an interest in the subject and may find it useful to pursue their studies. Basic concepts and proofs of theorems are explained in as lucid a manner as possible. Although the theory of probability is developed rigorously, it is developed in this book by a simple set theory approach. The statistical concepts are presented in an elegant manner. Unless the students become personally involved in solving exercises, they cannot really develop an understanding and appreciation of the ideas and a familiarity with the relevant techniques. Multiple choice questions are given at the end of each unit, with the answers also provided at the end. The solutions for all workbook exercises are given at the end of the book.

Organization of the Units

Unit 1

It deals with basic concepts of permutations, combinations, probability theory with some axioms, and theorems such as addition theorem. Some concepts such as conditional probability, multiplication theorem of probability, independent events, and Bayes' theorem with some of its applications are dealt with in this unit.

Unit 2

This unit describes discrete and continuous random variables with some of their properties given as probability mass functions and probability density functions. Also, joint probability distributions are dealt with in detail. Transformation of one- and two-dimensional random variables is found in this unit.

Unit 3

This unit introduces the basic concepts of expectation, variance, co-variance, and conditional expectation. Chebychev's theorem is given along with the proof. Central moments and raw moments are described. Moment-generating function and characteristic function are described along with their properties.

Unit 4

This unit concentrates on standard discrete distributions such as binomial distributions, Poisson distributions, negative binomial distributions, hypergeometric distributions, and geometric and uniform distributions. Some of the properties such as additive property, mean, variance, and moment-generating functions for each of the above distributions are described.

Unit 5
This unit focuses on continuous distributions such as normal distribution, exponential distribution, gamma distribution, and Weibull distributions. Many properties of normal distribution along with its importance and fitting are discussed. The moment-generating function for each of the above distributions is discussed here. Memory loss property of exponential distribution, failure rate of exponential distributions, and central limit theorem are described.

Unit 6
This unit outlines the definitions of sampling theory, its advantages, and its different types. Sampling distribution of means and difference between the means and variance are discussed.

Unit 7
This unit explains basic definitions in tests of significance for large samples, the steps involved in it, and the errors in sampling. Various tests of significance of single mean, difference of means, single proportion, difference of proportions, and difference of two standard deviations are discussed.

Unit 8
This unit deals with tests of significance for small samples such as the *t*-test for single mean and paired test for difference of means, *F*-test, and chi-square test. Chi-square test for goodness of fit and test for independence of attributes are discussed.

Unit 9
This unit elucidates the basic concepts of estimation such as point estimation and interval estimation. The characteristics of a good estimator are described. The concept of confidence interval and limits are given for single mean, proportion, difference between two means, difference between two proportions, variance, and ratio of two variances.

Unit 10
This unit presents the method of least squares for fitting a straight line and fitting various curves such as second-degree parabola, power curves of different types, and exponential curves.

Unit 11
This unit introduces correlation and its different types, methods, and properties. It also deals with coefficient of correlation for grouped data, its properties, rank correlation, limitations of Spearman's correlation coefficient, and the method to calculate correlation coefficient when ranks are tied.

Unit 12
This unit explains the concept of regression, lines of regression, and why there are two lines of regression. It also deals with regression coefficients, its properties, and the main differences between correlation and regression analysis. The angles between two lines of regression, standard error of estimate, and regression curves are also discussed.

Unit 13
This unit introduces basic definitions of queuing theory along with distribution of interarrival time, distribution of service time, and queuing process. It also describes different queuing models such as single-server and multiserver models. Five models and their results are discussed with several examples.

Unit 14
This unit focuses on the assumptions of analysis of variance and its different types, such as one-way classification and two-way classification. Other concepts such as completely randomized design, Latin square design, and randomized block design are also dealt with examples.

Unit 15

This unit expounds random process, discrete random process, continuous random process, deterministic and nondeterministic random process, stationarity, and distribution and density functions. First-order and second-order stationarity processes and wide-sense stationarity process are described. Other concepts such as cross-correlation function, statistical and time averages, ergodic random process, mean ergodic theorem, correlation ergodic process and correlation functions, and autocorrelation and cross-correlation function with their properties are discussed.

Unit 16

This unit introduces Poisson process and its probability law, second-order probability function of a homogeneous Poisson process, mean and autocorrelation of Poisson process, Markov process, and Chapman–Kolmogorov theorem. Some definitions in Markov chain such as irreducible Markov chain and irreversible periods are given. Gaussian process and its properties, narrow-band Gaussian process, and band-limited process are discussed.

Acknowledgements

I am grateful to Dr V. Ravindranath, Professor and Head, Department of Mathematics, JNTUK, Kakinada, Andhra Pradesh, India, for helping me to give a shape to this book. I also express my thanks to Sri. K. Anjaneya Murthy, Professor, Department of ECE, Vignana Bharathi Institute of Technology, Hyderabad, Andhra Pradesh, India, for helping me with the last units.

I wish to express my thanks to Dr N. Goutham Rao, Chairman, Dr J. S. N. Murthy, Principal, Dr Jayant Kulkarni, Vice Principal, Dr C. R. N. Sarma, Head of the Department of Humanities and Sciences, Vignana Bharathi Institute of Technology, Hyderabad, for the support extended during this project.

I wish to express my heartfelt thanks to my husband, Er C. Subrahmanyam and my children Mr C. Ravichandran and Mr C. V. S. K. Krishnan for their continuous support throughout this project. My special thanks are due to my father Sri P. V. Ramachandra Rao, who is the motivational factor for this project.

I am thankful to the publishers, Pearson Education, Chennai, for their continuous efforts and cooperation in bringing out this book within a short span of time.

I am sure that the students and the faculty will find this book very useful. Critical evaluation and suggestions for improvement of the book will be highly appreciated and gratefully acknowledged.

P. Kousalya

This page is intentionally left blank.

1 Probability

Prerequisites

Before you start reading this unit, you should:

- Know elementary concepts of set theory
- Have knowledge in permutations and combinations

Learning Objectives

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

- Understand the concepts of probability and conditional probability
- Understand addition theorem and multiplication theorems
- Understand Baye's theorem
- Apply the concept of Baye's theorem to some numerical examples

INTRODUCTION

The Unit starts with some elementary concepts of set theory, basic definitions required for probability, and the axioms of probability. It also covers the definition of conditional probability, multiplication theorem, and some theorems based on independent events. The Unit ends with Baye's theorem and problems on it.

1.1 ELEMENTARY CONCEPTS OF SET THEORY

If A, B and C are any three sets then

- *Commutative laws:*

$$A \cup B = B \cup A$$

$$A \cap B = B \cap A$$

- *Associative laws:*

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

$$A \cap (B \cap C) = (A \cap B) \cap C$$

- *Distributive laws:*

$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$$

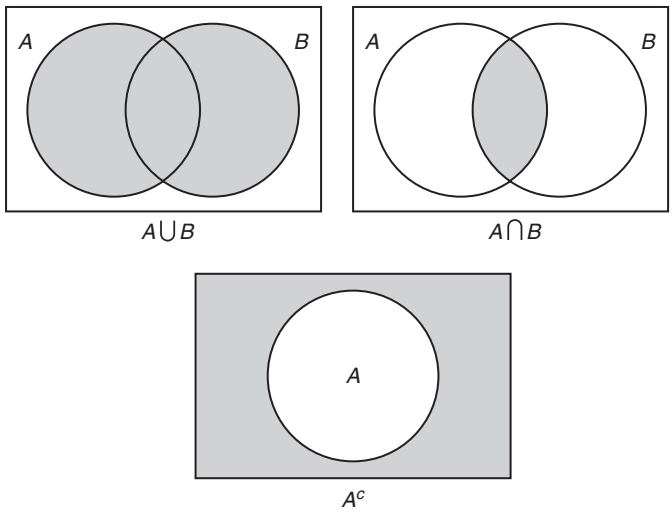
$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$$

- *De Morgan's laws:*

$$(A \cup B)^c = A^c \cap B^c$$

$$(A \cap B)^c = A^c \cup B^c$$

- Some concepts through Venn diagrams



1.2 PERMUTATIONS AND COMBINATIONS

The number of ways of arranging r things at a time from n available things is called permutation. It is denoted by nP_r.

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

The number of ways of selecting r things at a time from r available things is called as combination. It is denoted by nC_r.

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

Caution:

- The order of arrangements (permutations) is important whereas the order of selections (combinations) is not important.
- $nC_r = \frac{nP_r}{r!}$.

1. **Product rule:** If a first task can be performed in n_1 ways and second task can be performed in n_2 ways then the number of ways both the tasks can be performed is $n_1 \times n_2$ ways.
2. **Sum rule:** If a first task can be performed in n_1 ways and the second task can be performed in n_2 ways, then the number of ways of doing either of the tasks is $n_1 + n_2$ ways.

Now, let us apply the concepts developed so far in working out some problems.

Worked Out Examples

EXAMPLE 1.1

How many baseball teams each containing nine members can be chosen from among 12 boys without regard to the position played by each member?

Solution: Since there is no regard to the position, this is a combination.

Number of ways of choosing 9 members at a time from 12 members

$$12C_3 = \frac{12 \cdot 11 \cdot 10}{1 \cdot 2 \cdot 3}$$
$$= 220 \text{ ways.}$$

EXAMPLE 1.2

In how many ways can the director of a research laboratory choose two chemists from among seven applicants and three physicists from among nine applicants?

Solution: Number of ways of choosing 2 chemists from 7 applicants

$$7C_2 = \frac{7 \cdot 6}{1 \cdot 2} = 21 \text{ ways}$$

Number of ways of choosing 3 physicists from 9 applicants

$$9C_3 = \frac{9 \cdot 8 \cdot 7}{1 \cdot 2 \cdot 3} = 84 \text{ ways}$$

Using product rule, the director of research laboratory can chose 2 chemists and 3 physicists

$$= 21 \times 84$$
$$= 1365 \text{ ways.}$$

EXAMPLE 1.3

A carton of 12 rechargeable batteries contains one that is defective. In how many ways can an inspector choose three of the batteries and

(i) Get the one that is defective

(ii) Do not get the one that is defective?

Solution:

(i) Since one defective battery is to be chosen from one defective battery, it can be done in 1 way. The remaining two non-defective batteries can be chosen from 11 non-defective batteries in the following method:

$$11C_2 = \frac{11 \cdot 10}{1 \cdot 2}$$
$$= 55 \text{ ways}$$

Hence, the inspector can make a selection by getting the battery that is defective in 55 ways.

(ii) Since there should be no defective batteries in the selection, keep aside the defective one. Hence 3 non-defective batteries can be chosen from among 11 non-defective ones in the following method:

$$11C_3 = \frac{11 \cdot 10 \cdot 9}{1 \cdot 2 \cdot 3}$$

$$= 165 \text{ ways}$$

Hence, the inspector can make a selection by not getting the defective one in 165 ways.

Work Book Exercises

1. Consider Example 1.3. In this problem suppose two of the batteries are defective then in how many ways can the inspector choose 3 of the batteries and get
   (i) None of the defective batteries
   (ii) One of the defective batteries
   (iii) Both of the defective batteries?
2. In how many ways can the letters of the word 'ABACUS' be rearranged such that the vowels always appear together?
3. How many different four letter words can be formed (the words need not be meaningful) using the letters of the word 'MEDITERRANEAN' such that the first letter is *E* and the last letter is *R*?
4. In how many ways can the letters of the word 'PROBLEM' be rearranged to make seven letter words such that none of the letters repeat?

1.3 INTRODUCTION OF PROBABILITY

Probability is a word which we come across in our day-to-day life. A teacher may expect better results from section-A than section-B. This expectation comes from previous knowledge. We need a quantitative measure for these expectations, which are by our previous experience and present knowledge. Hence, the theory of probability plays a very important role. More mathematically the definition can be drawn as follows:

If an experiment is repeated under essentially identical conditions, generally we come across the following two situations:

1. The result of the experiment which is called as outcome is not known, or can be predicted or which is one from the several possible outcomes.
2. The outcome of the experiment is unique.

In the first case it is deterministic, and in the second case it is probabilistic. The second case phenomena are observed in our daily life.

For example, according to the Boyles law, {gas law}, PV = constant for a perfect gas where, P is the pressure of the gas and V is the volume of the gas. This expression is deterministic, where as a coin is thrown, its outcome cannot be predicted. Hence, our interest is to know the chances (or probability) of such throws, where we predict the result with the help of the previous knowledge. So here is how the term 'probability' can be defined with the help of some other terms given below.

Some Definitions of Probability

1. **Random experiment:** An experiment whose outcome or results are not unique and which cannot be predicted with certainty is called random experiment.

   Example: Tossing of a coin, throwing a dice, etc.

2. **Sample space:** The set of all outcomes of a random experiment is called sample space and is denoted by S.

   Example: When a dice is thrown the sample space is, $S = \{1, 2, 3, 4, 5, 6\}$.

3. **Event:** It is a subset of sample space.

   Example: $E_1 = \{1, 3, 5\}$.

   The UNION of two events A and B denoted by $A \cup B$ is the event containing all the elements that belong to A or B or both.

   Example: Let $A = \{1, 2, 3\}$, $B = \{3, 4, 5, 6\}$

   Then, $A \cup B = \{1, 2, 3, 4, 5, 6\}$

   The INTERSECTION of two events A and B denoted by $A \cap B$ is the event containing all the elements that are common to A and B.

   Example: Let $A = \{1, 2, 3\}$ and $B = \{4, 5, 6, 7\}$

   Then, $A \cap B = \{\}$

   The COMPLEMENT of an event A with respect to sample S is the subset of all the elements of S that are not in A. It is denoted by A^c or A'.

   Results:

   (i) $A \cap \emptyset = \emptyset$

   (ii) $A \cup \emptyset = A$

   (iii) $A \cap A^c = \emptyset$

   (iv) $A \cup A^c = S$

   (v) $\emptyset^c = S$

   (vi) $S^c = \emptyset$

   (vii) $(A^c)^c = A$.

4. **Mutually exclusive events:** Two events A and B are said to be mutually exclusive events if the occurrence of one event excludes (precludes) the occurrence of the other event.

   Example: When a coin is thrown, the occurrence of head excludes the occurrence of tail.

   Caution: A and B are mutually exclusive events if $A \cap B = \emptyset$ (A and B are disjoint sets).

5. **Collectively exhaustive events:** A list of events $A_1, A_2, A_3, A_4, \ldots A_n$ are collectively exhaustive if the union of all the events is the entire sample space.

   Caution: $A_1, A_2, A_3, A_4, \ldots A_n$ are collectively exhaustive if $\bigcup_{i-1}^{n} A_i = S$

6. **Equally likely events:** The events A and B are said to be equally likely events if each of the elements have equal chance of occurrence.

Mathematical or Classical Probability

If an event E can occur in m ways out of n mutually exclusive, equally likely and collectively exhaustive ways, then the probability of occurrence of an event E, denoted by $P(E)$ is given by

$$P(E) = \frac{\text{Favourable number of cases for } E}{\text{Total number of cases}}$$

$$P(E) = \frac{m}{n}$$

Caution:

- The probability of occurrence of an event E is called its success, denoted by $P(E)$ and its non-occurrence is called its failure, denoted by $P(\bar{E})$.
- The probability of non-occurrence of the event is given as follows:

$$P(\bar{E}) = \frac{\text{Favourable number of cases for } \bar{E}}{\text{Total number of cases}}$$

$$= \frac{n - m}{n}$$

$$= 1 - \frac{m}{n}$$

$$P(\bar{E}) = 1 - P(E)$$

$$\therefore P(E) + P(\bar{E}) = 1$$

$$\therefore (p + q) = 1$$

- Probability of a sure event (certain event) is always 1.
- Probability of an impossible event is 0.
- The definition of probability fails when the outcomes are not equally likely.
- When the outcomes are infinite, probability of occurrence of an event E is given by $P(E) = \underset{n \to \alpha}{Lt} \frac{m}{n}$, where m is the number of times an event E happens in n trials assuming that the trials are performed under essentially homogeneous and identical conditions.

1.4 AXIOMS OF PROBABILITY

The three axioms of probability are as follows:

1. For any event E of S,

$$0 \le P(E) \le 1$$

$\therefore$ Probability of an event is a numerical value, which always lies between 0 and 1.

2. Probability of entire sample space is always 1.
3. For any two mutually exclusive events E_1 and E_2 of S, probability of occurrence of either E_1 or E_2 is given by,

$$P(E_1 \cup E_2) = P(E_1) + P(E_2)$$

1.5 SOME ELEMENTARY RESULTS

RESULT 1: If Ø is null set then $P(Ø) = 0$.

Proof: If S is the entire sample space then

$$S = S \cup Ø$$

$$P(S) = P(S \cup Ø)$$

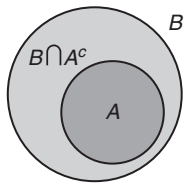
Since the sample space and null set are mutually exclusive events, from axiom (3) of probability,

$$P(S) = P(S) + P(Ø)$$

$$P(Ø) = 1.$$

RESULT 2: If $A \subseteq B$, then $P(A) \leq P(B)$.

Proof: From the diagram, the event B can be written as



$$B = A \cup (A^c \cap B)$$

$$P(B) = P[A \cup (A^c \cap B)]$$

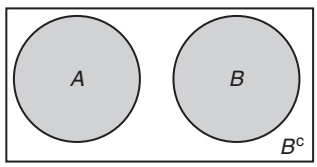
Since from the diagram it is clear that, A and $(A^c \cap B)$ are mutually exclusive events and hence from axiom (3) of probability,

$$P(B) = P(A) + (A^c \cap B)$$

However, $P(A^c \cap B)$ is never zero and a positive quantity, since $A \subseteq B$

$$P(B) \geq P(A)$$

$$\therefore P(A) \leq P(B).$$



RESULT 3: If $A \cap B = Ø$, then $P(A) \leq P(B^c)$.

Proof: Since $A \cap B = Ø$ we know that $A \subseteq B^c$
However, from the previous theorem it is clear that when $A \subseteq B^c$
Then, $P(A) \leq P(B^c)$.

RESULT 4: If A is any event of S then $0 \leq P(E) \leq 1$.

Proof: Since $Ø \subseteq A$ and from Result 2,

$$P(Ø) \leq P(A) \quad \{\because P(Ø) = 0\}$$

$$0 \leq P(A) \tag{1.1}$$

In addition, since A which is any set of S, the sample space

$$A \subseteq S$$

$$P(A) \le P(S) \quad \{\because P(S) = 1\}$$

$$P(A) = 1 \tag{1.2}$$

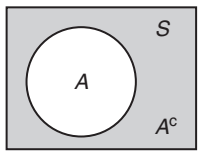
From (1.1) and (1.2) we have

$$0 \le P(E) \le 1.$$

RESULT 5: If A is any event in S, then $P(A^c) = 1 - P(A)$.

Proof: We know that $A \cup A^c = S$

$$P(A \cup A^c) = P(S) = 1 \tag{1.3}$$



Since A and A^c are mutually exclusive events, from axiom (3) of probability

$$P(A \cup A^c) = P(A) + P(A^c)$$

Hence from (1.3),

$$P(A) + P(A^c) = 1$$

$$\therefore P(A^c) = 1 - P(A)$$

Caution:

- The odds in favour of an event A is $P(A):P(A^c)$.
- The odds against an event A is $P(A^c):P(A)$.

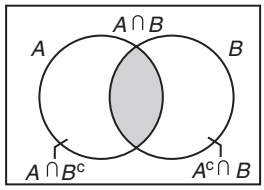
RESULT 6: For any two events A and B,

$$P(A^c \cap B) = P(B) - P(A \cap B)$$

Proof: It is clear from the Venn diagrams, that

$$B = (A^c \cap B) \cup (A \cap B)$$

$$P(B) = P[(A^c \cap B) \cup (A \cap B)]$$



Since $(A^c \cap B)$ and $(A \cap B)$ are mutually exclusive events, from axiom (3) of probability,

$$P(B) = P(A^c \cap B) + P(A \cap B)$$

$$P(A^c \cap B) = P(B) - P(A \cap B)$$

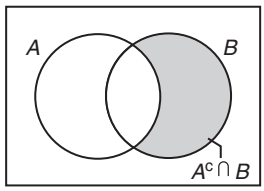
Caution: From the above Venn diagrams, we get

$$A = (A \cap B^c) \cup (A \cap B)$$

$$\therefore P(A) = P(A \cap B^c) + P(A \cap B).$$

RESULT 7: Addition theorem of probability—for any two events, A and B of S, $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

Proof: From the Venn diagram it has been observed that



$A \cup B$ can be written as the union of two mutually exclusive events, A and $(A^c \cap B)$.

$$A \cup B = A \cup (A^c \cap B)$$

$$P(A \cup B) = P[A \cup (A^c \cap B)]$$

From axiom (3) of probability, since A and $(A^c \cap B)$ are mutually exclusive events,

$$P[A \cup (A^c \cap B)] = P(A) + P(A^c \cap B)$$

$$\therefore P(A \cup B) = P(A) + P(A^c \cap B) \tag{1.4}$$

However, from Result 6, $P(A^c \cap B) = P(B) - P(A \cap B)$

Substituting this in equation (1.4), we get

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

Caution: If A and B are mutually exclusive events $P(A \cap B) = 0$
Hence, from the above theorem, we get

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

RESULT 8: For any three events A, B, and C, $P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) - P(A \cap B \cap C)$.

Proof: Take a substitution in the result to be proved as follows:

Let $$B \cup C = D \tag{1.5}$$

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

From Result 7, we get

$$P(A \cup D) = P(A) + P(D) - P(A \cap D) \tag{1.6}$$

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

From the distributive law, we get

$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$$

$$\begin{aligned} P(A \cap D) &= P[A \cap (B \cup C)] \\ &= P[(A \cap B) \cup (A \cap C)] \\ &= P(A \cap B) + P(A \cap C) - P[(A \cap B) \cap (A \cap C)] \\ &= P(A \cap B) + P(A \cap C) - P(A \cap B \cap C) \end{aligned}$$

{from Associative law of events of A, B, and C}

Substituting the above equation and equation (1.5) in equation (1.6), we have

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

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

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

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

Hence the result.
Now, let us apply the concepts developed so far in working out some problems.

Worked Out Examples

EXAMPLE 1.4

(i) A bag contains 4 white and 6 black balls. If a ball is drawn at random, find the probability for it to be black.

(ii) A bag contains 6 white, 8 black, and 5 red balls. Three balls are drawn from it at random. Find the probability that all the 3 balls are white.

Solution:

(i) Let E be the event of drawing a black ball.

$$\text{Probability for event } E = \frac{\text{Favourable number of cases for } E}{\text{Total number of cases}}$$

Favourable number of cases for E = number of ways of drawing one black ball out of 6 black balls

$$= 6C_1 = 6 \text{ ways}$$

Total number of cases = number of ways of drawing one ball out of a total of (4 + 6)

$$= 10 \text{ balls}$$
$$= 10C_1 = 10 \text{ ways}$$

$$\therefore\ P(E) = \frac{6}{10} = \frac{3}{5}.$$

(ii) Let E be the event of drawing 3 white balls.

$$\text{Probability of event } E = \frac{\text{Favourable number of cases for } E}{\text{Total number of cases}}$$

Favourable number of cases = number of ways of drawing 3 white balls from 6 white balls

$$= 6C_3 = \frac{6 \cdot 5 \cdot 4}{1 \cdot 2 \cdot 3}$$
$$= 20 \text{ ways}$$

Total number of cases = number of ways of drawing 3 balls from a total of (6 + 8 + 5) = 19 balls

$$= 19C_3 = \frac{19 \cdot 18 \cdot 17}{1 \cdot 2 \cdot 3} = 969 \text{ ways}$$

$$\therefore\ P(E) = \frac{20}{969}.$$

EXAMPLE 1.5

A book containing 100 pages is opened at random. Find the probability that on the page

(i) A doublet is found

(ii) A number whose sum of the digits is 10.

Solution:

(i) The doublets on the page numbers are = {11, 22, 33, 44, 55, 66, 77, 88, 99}.

Let E be the event of getting a page whose number is a doublet.

$$\text{Probability of } E = \frac{\text{Favourable number of cases for } E}{\text{Total number of cases}}$$

Favourable number of cases = drawing a number from the above list of numbers

= drawing one number from 9 numbers

$= 9C_1 = 9$ ways

Total number of ways = drawing one from the total of 100 pages

$= 100C_1 = 100$ ways

$$\therefore\ P(E) = \frac{9}{100}.$$

(ii) The numbers whose sum of the digits is 10 = {(1,9), (2,8), (3,7), (4,6), (5,5), (6,4), (7,3), (8,2), (9,1)}.

Let E be the event drawing a number whose sum of the digits is 10.

$$\text{Probability of } E = \frac{\text{Favourable number of cases for } E}{\text{Total number of cases}}$$

Favourable number of cases = number of ways of choosing one number from the above list

= number of ways of drawing one from the 9 pairs available

$= 9C_1 = 9$ ways

Total number of ways = drawing one number from 100 numbers

$= 100C_1 = 100$ ways

$$\therefore\ P(E) = \frac{9}{100}.$$

EXAMPLE 1.6

If $P(A) = \frac{1}{2}$, $P(B) = \frac{1}{3}$, and $P(A \cap B) = \frac{1}{5}$, then find the following:

(i) $P(A \cup B)$

(ii) $P(A^c \cap B)$

(iii) $P(A \cap B^c)$

(iv) $P(A^c \cap B^c)$ (JNTU Aug./Sep. 2007, set-1, Nov. 2006, set-4)

Solution:

(i) From addition theorem of probability,

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$
$$= \frac{1}{2} + \frac{1}{3} - \frac{1}{5}$$
$$= 0.633333$$

(ii) $P(A^c \cap B) = P(B) - P(A \cap B)$
$$= \frac{1}{3} - \frac{1}{5}$$
$$= 0.133333$$

(iii) $(A \cap B^c) = P(A) - P(A \cap B)$
$$= \frac{1}{2} - \frac{1}{5}$$
$$= 0.3$$

(iv) $A^c \cap B^c = (A \cup B)^c$

$$P(A^c \cap B^c) = P(A \cup B)^c$$
$$P(A^c \cap B^c) = 1 - P(A \cup B)$$
$$= 1 - \frac{1}{5} = 0.8.$$

EXAMPLE 1.7

If $P(A) = a$, $P(B) = b$, $P(A \cap B) = c$, express the following in terms of a, b and c.

(i) $P(A^c \cup B^c)$
(ii) $P(A \cup B^c)$
(iii) $P(A^c \cap B)$
(iv) $P(A^c \cap B^c)$
(v) $P(A^c \cup B)$
(vi) $P[A^c \cap (A \cup B)]$
(vii) $P[A \cup (A^c \cap B)]$

Solution:

(i) From De Morgan's laws,

$$A^c \cup B^c = (A \cap B)^c$$
$$P(A^c \cup B^c) = P(A \cap B)^c$$
$$= 1 - P(A \cap B)$$
$$= 1 - c$$

(ii) $P(A \cap B^c) = P(A) - P(A \cap B)$
$$= a - c$$

(iii) $P(A^c \cap B) = P(B) - P(A \cap B)$
$$= b - c$$

(iv) From De Morgan's laws,

$$A^c \cap B^c = (A \cup B)^c$$

$$P(A^c \cap B^c) = P(A \cup B)^c$$

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

$$= 1 - [P(A) + P(B) - P(A \cap B)]$$

$$= 1 - [a + b - c]$$

$$= 1 - a - b + c$$

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

$$= 1 - P(A) + P(B) - P(A^c \cap B)$$

$$= 1 - a + b - (b - c)$$

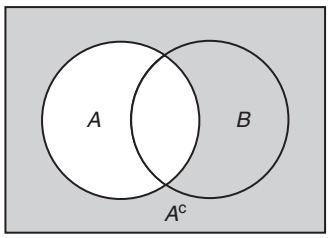
$$= 1 - a + b - b + c$$

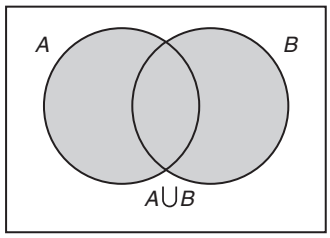
$$= 1 - a + c$$

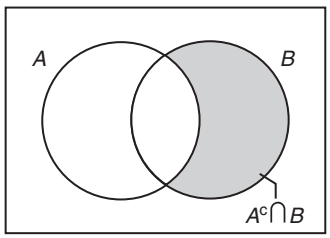
(vi) $A^c \cap (A \cup B) = A^c \cap B$

$$P[A^c \cap (A \cup B)] = P(A^c \cap B)$$

$$= b - c$$







(vii) $P[A \cup (A^c \cap B)] = P(A) + P(A^c \cap B) - P[A \cap (A^c \cap B)]$

$$= P(A) + P(A^c \cap B) - P(\emptyset)$$

$$= P(A) + P(B) - P(A \cap B) - 0$$

$$= a + b - c.$$

EXAMPLE 1.8

Five persons in a group of 25 are teachers. If 3 persons are selected at random, determine the probability that

(i) All are teachers

(ii) At least one is a teacher.

Solution:

(i) Let E be the event that there are 3 teachers among 5 teachers of a group of 25.

$$\text{Probability of } E = \frac{\text{Favourable number of cases for } E}{\text{Total number of cases}}$$

Favourable number of cases for E = number of ways of selecting three teachers from 5

$$= 5C_3 = 5C_2$$

$$= \frac{5 \cdot 4}{1 \cdot 2} = 10 \text{ ways}$$

Total number of cases = number of ways of selecting 3 teachers from a group of 25

$$= 25C_3 = \frac{25 \cdot 24}{1 \cdot 2 \cdot 3} = 100 \text{ ways}$$

$$\therefore P(E) = \frac{10}{100} = \frac{1}{10}.$$

(ii) Let E be the event that there is no teacher among the 5 teachers.

Probability that there is at least one teacher = 1 – Probability of 'no teacher' among the 5

$$\text{Probability of } E = \frac{\text{Favourable number of cases for } E}{\text{Total number of cases}}$$

Favourable number of cases = number of ways of selecting no teacher

= number of ways of selecting 3 non-teachers from the group of $(25 - 5) = 20$

$$= 20C_3$$

$$= \frac{20 \cdot 19 \cdot 18}{1 \cdot 2 \cdot 3} = 1140 \text{ ways}$$

Total number of cases = number of ways of selecting 3 from among 25

$$= 20C_3$$

$$= \frac{25 \cdot 24 \cdot 23}{1 \cdot 2 \cdot 3} = 4600 \text{ ways}$$

$$P(E) = \frac{1140}{4600}$$

$$= \frac{57}{230}$$

Probability of at least one teacher = $1 - P(E)$

$$= 1 - \frac{57}{230}$$

$$\text{Probability of at least one teacher} = \frac{173}{230}.$$

Caution: P(occurrence of atleast one event) = 1 – P(occurrence of none of the events).

EXAMPLE 1.9

Two marbles are drawn in succession from an urn containing 10 red, 30 white, 20 blue, and 15 orange marbles

(a) With replacement being made after each drawing. Find the probability that

(i) Both are white

(ii) First is red and second is white

(iii) Neither is orange

(b) When no replacement is made, solve the above problem.

Solution:

(a) When replacement is made:

(i) Let E_1 be the event of drawing a white and E_2 be the event of drawing a white after the first ball is replaced.

Probability that both marbles drawn are white = P (first is white and second is white)

$= P$ (first is white) $\cdot$ P(second is white)

$$= \frac{\text{Favourable number of cases for } E_1}{\text{Total number of cases}} \cdot \frac{\text{Favourable number of cases for } E_2}{\text{Total number of cases}}$$

Favourable number of cases for E_1 = Drawing one from 30 white balls

$= 30C_1 = 30$ ways

Total cases for E_1 = Drawing one from a total of $(10 + 30 + 20 + 15) = 75$ balls

$= 75C_1 = 75$ ways

$$P(E_1) = \frac{30}{75}$$

Favourable number of cases for E_2 = Drawing one from 30 white balls after replacement

$= 30C_1 = 30$ ways

Total cases for E_2 = Drawing one from a total of $(10 + 30 + 20 + 15) = 75$ balls

$= 75C_1 = 75$ ways

$$P(E_1) = \frac{30}{75}$$

Probability that both marbles drawn are white = $P(E_1) \cdot P(E_2)$

$$= \frac{30}{75} \cdot \frac{30}{75}$$

$$= \frac{4}{25}.$$

(ii) Let E_1 be the event of drawing a red marble and E_2 be the event of drawing a white marble after the first marble is replaced.

Probability that both marbles drawn are white = P(first is red and second is white)

$= P$(first is red) $\cdot$ P(second is white)

$$= P(E_1) \cdot P(E_2)$$
$$= \frac{10C_1}{75C_1} \cdot \frac{30C_1}{75C_1}$$
$$= \frac{10}{75} \cdot \frac{10}{75}$$
$$= \frac{4}{75}.$$

(iii) Let E_1 be the event of drawing not orange and E_2 be the event of drawing not orange after the first marble is replaced.

Probability that both marbles drawn are not orange = P(first is not orange and second is not orange)

$$= P(\text{first is not orange}) \cdot P(\text{second is not orange})$$
$$= \frac{60C_1}{75C_1} \cdot \frac{60C_1}{75C_1}$$
$$= \frac{60}{75} \cdot \frac{60}{75}$$

Probability that both the marbles drawn are not orange = $\frac{16}{25}$.

(b) When replacement is not made

(i) Let E_1 be the event of drawing a white and E_2 be the event of drawing a white after the first ball is not replaced.

Probability that both marbles drawn are white = P(first is white and second is white)

$$= P(\text{first is white}) \cdot P(\text{second is white})$$
$$= \frac{30C_1}{75C_1} \cdot \frac{29C_1}{74C_1}$$
$$= \frac{30}{75} \cdot \frac{29}{75}$$
$$= \frac{58}{375}.$$

EXAMPLE 1.10

Two dice, one green and the other red are thrown. Let A be the event that the sum of the faces shown on the dice is odd. Let B be the event of at least one ace (number 1) on the faces of the dice.

(i) Describe the complete sample space

(ii) Describe events A, B, B^c, $A \cap B$, $A \cup B$, $A \cap B^c$

(iii) Find $P(A^c \cup B^c)$, $P(A^c \cap B^c)$, $P(A \cap B^c)$, $P(A^c \cap B)$

(iv) Find $P[A^c \cap (A \cup B)]$

(v) Find $P\left(\frac{A}{B}\right)$, $P\left(\frac{B}{A}\right)$, $P\left(\frac{A^C}{B^C}\right)$, $P\left(\frac{B^C}{A^C}\right)$

Solution:

(i) The complete sample space is as follows:

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

$n(S)$ = Total number of elements in sample space = 36

(ii) Let A be the event that the sum of the faces shown on the dice is odd.

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

$n(A)$ = Total number of elements in A = 18.

$$P(A) = \frac{n(A)}{n(S)}$$
$$= \frac{18}{36}$$
$$= \frac{1}{2}$$

Let B be the event of at least one ace (number 1) on the faces of the dice.

B = {(1,1), (1,2), (1,3), (1,4), (1,5), (1,6), (2,1), (3,1), (4,1), (5,1), (6,1)}

$n(B)$ = Total number of elements in B = 11

$$P(A) = \frac{n(B)}{n(S)}$$
$$= \frac{11}{36}$$

B^c = {(2,2), (2,3), (2,4), (2,5), (2,6), (3,2), (3,3), (3,4), (3,5), (3,6), (4,2), (4,3), (4,4), (4,5), (4,6), (5,2), (5,3), (5,4), (5,5), (5,6), (6,2), (6,3), (6,4), (6,5), (6,6)}

$n(B^c) = 25$

$$P(B^c) = \frac{n(B)}{n(S)} = \frac{25}{36}$$
$$= 1 - P(B) = 1 - \frac{11}{36}$$
$$= \frac{25}{36}$$
$$= 0.694444$$

$A \cap B$ = {(1,2), (1,4), (1,6), (2,1), (4,1), (6,1)}

$n(A \cap B) = 6$

$$P(A \cap B) = \frac{6}{36} = \frac{1}{6}$$

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

$$= \frac{1}{2} + \frac{11}{36} - \frac{1}{6}$$

$$= 0.638889$$

(iii) $$P(A^c \cup B^c) = 1 - P(A \cap B)$$

$$= 1 - \frac{1}{6}$$

$$= 0.833333$$

$$P(A^c \cap B^c) = 1 - P(A \cup B)$$

$$= 1 - 0.638889$$

$$= 0.361111$$

$$P(A^c \cup B^c) = 1 - P(A \cap B)$$

$$= 1 - \frac{1}{6}$$

$$= 0.833333$$

$$P(A^c \cap B) = P(B) - P(A \cap B)$$

$$= \left(\frac{11}{36}\right) - \left(\frac{1}{6}\right)$$

$$= 0.138889$$

$$P(A \cap B^c) = P(A) - P(A \cap B)$$

$$= \left(\frac{1}{2}\right) - \left(\frac{1}{6}\right)$$

$$= 0.333333$$

(iv) $$A^c \cap (A \cup B) = (A^c \cap A) \cup (A^c \cap B)$$

$$P\{A^c \cap (A \cup B)\} = P\{(A^c \cap A) \cup (A^c \cap B)\}$$

$$= P\{\cup (A^c \cap B)\}$$

$$= P\{(A^c \cap B)\}$$

$$= 0.138889$$

(v) $$P\left(\frac{A}{B}\right) = \frac{P(A \cap B)}{P(B)}$$

$$= \frac{\frac{1}{6}}{\frac{11}{36}}$$

$$= \frac{6}{11}$$

$$= 0.545455$$

$$P\left(\frac{B}{A}\right) = \frac{P(A \cap B)}{P(A)}$$

$$P\left(\frac{B}{A}\right) = \frac{\frac{1}{6}}{\frac{18}{36}}$$

$$= \frac{1}{3} = 0.33333$$

$$P\left(\frac{A^c}{B^c}\right) = \frac{P(A^c \cap B^c)}{P(B^c)}$$

$$= \frac{0.361111}{0.694444}$$

$$= 0.52$$

$$P\left(\frac{B^c}{A^c}\right) = \frac{P(A^c \cap B^c)}{P(A^c)}$$

$$= \frac{0.361111}{0.5}$$

$$= 0.722222$$

Work Book Exercises

5. A four digit number is formed with the digits 1, 2, 3, 4, and 5 with no repetitions of any digit. Find the chance that the number is divisible by 5.
6. By comparing appropriate regions of Venn diagrams, verify that
   (i) $(A \cap B) \cup (A \cap B^c) = A$
   (ii) $A^c \cap (B^c \cup C) = (A^c \cap B^c) \cup (A^c \cap C)$
7. Two cards are selected at random from 10 cards numbered 1 to 10. Find the probability that the sum is even if,
   (i) The two cards are drawn together
   (ii) The two cards are drawn one after the other with replacement.

(JNTU Aug./Sep. 2008, set-4)

1.6 CONDITIONAL PROBABILITY

The probability of an event occurring under a condition is called conditional probability. If an event A has already occurred, then the probability of another event B with the condition that A has already occurred is called the conditional probability of B and is denoted by $P\left(\frac{B}{A}\right)$. It is given by the following equation:

$$P\left(\frac{B}{A}\right) = \frac{P(A \cap B)}{P(A)}$$

Similarly, $P\left(\frac{A}{B}\right) = \frac{P(A \cap B)}{P(B)}$

The following theorem gives the probabilities of the simultaneous occurrences of the two events A and B:

Multiplication Theorem of Probability

For any two events A and B,

$$P(A \cap B) = P(A) \cdot P\left(\frac{B}{A}\right), \quad P(A) > 0$$

$$= P(B) \cdot P\left(\frac{A}{B}\right), \quad P(B) > 0$$

Where, $P\left(\frac{B}{A}\right)$ represents the conditional probability of occurrence of B when the event A has already happened. $P\left(\frac{A}{B}\right)$ represents the conditional probability of occurrence of A when the event B has already happened.

Proof: We know that

$$P(A) = \frac{n(A)}{n(S)} \tag{1.7}$$

$$P(B) = \frac{n(B)}{n(S)} \tag{1.8}$$

$$P(A \cap B) = \frac{n(A \cap B)}{n(S)} \tag{1.9}$$

For the event $A \cap B$, the favourable outcomes must be one of the sample points of B, that is, for the event $\left(\frac{A}{B}\right)$, the sample space is B and out of $n(B)$ sample points, $n(A \cap B)$ are the occurrences of the event A.

Hence, $P\left(\frac{A}{B}\right) = \frac{n(A \cap B)}{n(B)}$ (1.10)

From equation (1.9),

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

$$= \left(\frac{n(B)}{n(S)}\right)\left(\frac{n(A \cap B)}{n(B)}\right)$$

$$P(A \cap B) = P(B) \cdot P\left(\frac{A}{B}\right)$$

Similarly $P(A \cap B) = P(A) \cdot P\left(\frac{B}{A}\right)$.

Independent Events

Two events are said to be independent if the occurrence of one event affects the occurrence of the other event. If $P\left(\frac{A}{B}\right) = P(A)$ or if $P\left(\frac{B}{A}\right) = P(B)$, then the two events A and B are independent events.

From multiplication theorem of probability,

$$P(A \cap B) = P(A) \cdot P(B)$$

Caution: Please observe the difference between mutually exclusive events and independent events.

Pairwise Independent Events

If A, B, and C are events of a sample space S, they are said to be pair wise independent if $P(A \cap B) = P(A) \cdot P(B)$, $P(B \cap C) = P(B) \cdot P(C)$, $P(A \cap C) = P(A) \cdot P(C)$ when $P(A) \neq 0$, $P(B) \neq 0$, and $P(C) \neq 0$.

RESULT 9: If A and B are two independent events then A and B^c are also independent events.

Proof: Given that A and B are independent events,

$$P(A \cap B) = P(A) \cdot P(B) \tag{1.11}$$

To prove that A and B^c are independent, we have to show that

$$P(A \cap B^c) = P(A) \cdot P(B^c) \tag{1.12}$$

Consider,

$$\begin{aligned} P(A \cap B^c) &= P(A) - P(A \cap B) \\ &= P(A) - P(A) \cdot P(B), \text{ from equation (1.11)} \\ &= P(A)[1 - P(B)] \\ &= P(A) \cdot P(B^c) \end{aligned}$$

Hence, from equation (1.12) A and B^c are also independent events.

RESULT 10: If A and B are two independent events then A^c and B are also independent events.

Proof: Given that A and B are independent events,

$$P(A \cap B) = P(A) \cdot P(B) \tag{1.13}$$

To prove that A^c and B are independent, we have to show that

$$P(A^c \cap B) = P(A^c) \cdot P(B) \tag{1.14}$$

Consider,

$$\begin{aligned} P(A^c \cap B) &= P(B) - P(A \cap B) \\ &= P(B) - P(A) \cdot P(B), \text{ from equation (1.13)} \\ &= P(B)[1 - P(A)] \\ &= P(A^c) \cdot P(B) \end{aligned}$$

Hence, from equation (1.13) A^c and B are also independent events.

RESULT 11: If A and B are two independent events then A^c and B^c are also independent events.

Proof: Given that A and B are independent events,

$$P(A \cap B) = P(A) \cdot P(B) \tag{1.11}$$

To prove that A^c and B^c are independent, we have to show that

$$P(A^c \cap B^c) = P(A^c) \cdot P(B^c) \tag{1.15}$$

Consider,

$$\begin{aligned} A^c \cap B^c &= (A \cup B)^c \\ P(A^c \cap B^c) &= P(A \cup B)^c \\ &= 1 - P(A \cup B) \\ &= 1 - [P(A) + P(B) - P(A \cap B)] \\ &= 1 - P(A) - P(B) + P(A \cap B) \\ &= 1 - P(A) - P(B) + P(A) \cdot P(B) \\ &= [1 - P(A)][1 - P(B)] \\ P(A^c \cap B^c) &= P(A^c) \cdot P(B^c) \end{aligned}$$

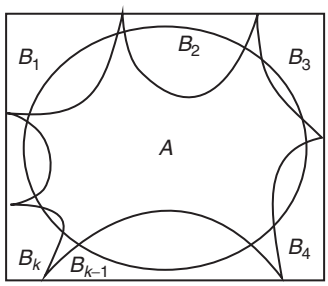
Hence, A^c and B^c are also independent events.

1.7 THEOREM OF TOTAL PROBABILITY

Let $B_1, B_2, B_3, \ldots B_k$ constitute a partition of the sample space S with $P(B_i) \neq 0$, for any $i = 1, 2, \ldots k$. Then for any event A of S,

$$\begin{aligned} P(A) &= \sum_{i=1}^{k} P(B_i \cap A) \\ &= \sum_{i=1}^{k} P(B_i) P\left(\frac{A}{B_i}\right) \end{aligned}$$

Proof: Since the entire sample space is partitioned into k subsets, their union is S and all B_i's are mutually disjoint sets.



Since $B_1, B_2, B_3, \ldots B_k$ constitute a partition,

$$S = \bigcup_{i=1}^{k} B_i \tag{1.16}$$

$$B_i \cap B_j = \varnothing \text{ for any } i \text{ and } j \tag{1.17}$$

Now, let A be any event which is constructed in such a way that $A \cap B_i$ for $i = 1, 2, \ldots k$ is not empty.

Now, we know that

$$\begin{aligned} A &= A \cap S \\ &= A \cap \left(\bigcup_{i=1}^{k} B_i \right) \text{ from equation (1.16)} \\ &= A \cap (B_1 \cup B_2 \cup B_3 \cup \ldots \cup B_k) \\ &= (A \cap B_1) \cup (A \cap B_2) \cup (A \cap B_3) \cup \ldots (A \cup B_k) \quad \{\text{from distributive law of events}\} \end{aligned}$$

From the above Venn diagram, it is clear that the events $A \cap B_1, A \cap B_2, \ldots A \cup B_k$ are mutually disjoint.

$$P(A) = P[(A \cap B_1) \cup (A \cap B_2) \cup (A \cap B_3) \cup \ldots (A \cup B_k)]$$

From axiom (3) of probability, since the events are mutually disjoint,

$$P(A) = P(A \cap B_1) + P(A \cap B_2) + \ldots + P(A \cup B_k)$$

$$= \sum_{i=1}^{k} P(A \cap B_i)$$

$$P(A) = \sum_{i=1}^{k} P(B_i) P(A/B_i) \tag{1.18}$$

1.8 BAYE'S THEOREM

Now, let us discuss about the important theorem of this unit which can be applied in certain situations that are given after the theorem and are proved. Suppose we come across situations with certain number of machines, we select a product randomly and find it to be defective. If we want to prove that this product comes from a particular machine say, Machine B_i then Baye's theorem can be used.

Let the sample space S be partitioned into k subsets $B_1, B_2, \ldots B_k$ with $P(B_i) \neq 0$.

for $i = 1, 2, \ldots k$. For any arbitrary event A in S, with $P(A) \neq 0$ and any specific event B_s, we have

$$P\left(\frac{B_s}{A}\right) = \frac{P(B_s \cap A)}{\sum_{i=1}^{n} P(B_i \cap A)}$$

$$= \frac{P(B_s) \cdot P\left(\frac{A}{B_s}\right)}{\sum_{i=1}^{k} P(B_i) \cdot P\left(\frac{A}{B_i}\right)} \quad \text{for } s = 1, 2, \ldots k$$

Proof: Recollect the definition of conditional probability, which says

$$P\left(\frac{B_s}{A}\right) = \frac{P(B_s \cap A)}{P(A)} \tag{1.19}$$

From the theorem of total probability,

$$P(A) = \sum_{i=1}^{k} P(B_i) \cdot P\left(\frac{A}{B_i}\right)$$

From multiplication theorem of probability,

$$P(B_s \cap A) = P(B_s) \cdot P\left(\frac{A}{B_s}\right) \tag{1.20}$$

Substituting equations (1.18) and (1.20) in (1.19) we get,

$$P\left(\frac{B_s}{A}\right) = \frac{P(B_s) \cdot P\left(\frac{A}{B_s}\right)}{\sum_{i=1}^{k} P(B_i) \cdot P\left(\frac{A}{B_i}\right)}$$

Now, let us apply the concepts developed so far in working out some problems.

Worked Out Examples

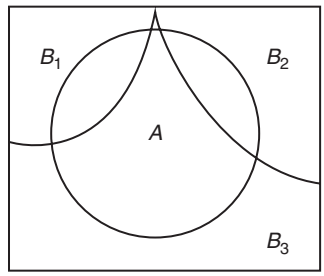
EXAMPLE 1.11

Companies B_1, B_2, and B_3 produce 30%, 45%, and 25% of the cars, respectively. It is known that 2%, 3%, and 2% of these cars produced from are defective.

(i) What is the probability that a car purchased is defective?

(ii) If a car purchased is found to be defective, what is the probability that this car is produced by B_1? (JNTU Aug./Sep. 2005)

Solution: Let B_1, B_2, and B_3 denote the events that the companies produce 30%, 45%, and 25% of the cars, respectively.



$$P(B_1) = 30\% = 0.3$$
$$P(B_2) = 45\% = 0.45$$
$$P(B_3) = 25\% = 0.25$$

Let A be an event that the cars produced are defective.

$P\left(\frac{A}{B_1}\right)$ = Probability that the cars produced by company B_1 has 2% of defectives

$= 0.02$

$P\left(\frac{A}{B_2}\right)$ = Probability that the cars produced by company B_2 has 3% of defectives

$= 0.03$

$P\left(\frac{A}{B_3}\right)$ = Probability that the cars produced by company B_3 has 2% of defectives

$= 0.02$

(i) The probability that a car purchased is defective = $P(A)$

From theorem of total probability we have,

$$P(A) = \sum_{i=1}^{k} P(B_i) \cdot P\left(\frac{A}{B_i}\right)$$

$$P(A) = P(B_1) \cdot P\left(\frac{A}{B_1}\right) + P(B_2) \cdot P\left(\frac{A}{B_2}\right) + P(B_3) \cdot P\left(\frac{A}{B_3}\right)$$

$$= (0.3)(0.02) + (0.45)(0.03) + (0.25)(0.02)$$

$$= 0.0245$$

(ii) Using Baye's theorem,

$$P\left(\frac{B_1}{A}\right) = \frac{P(B_1) \cdot P\left(\frac{A}{B_1}\right)}{P(B_1) \cdot P\left(\frac{A}{B_1}\right) + P(B_2) \cdot P\left(\frac{A}{B_2}\right) + P(B_3) \cdot P\left(\frac{A}{B_3}\right)}$$

$$= \frac{(0.3)(0.02)}{(0.03)(0.02) + (0.45)(0.03) + (0.25)(0.02)}$$

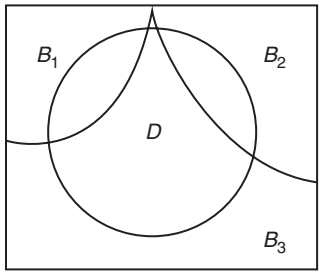
$$= 0.898876.$$

EXAMPLE 1.12

The probabilities of A, B, and C to become MD's of a factory are $\frac{5}{10}$, $\frac{3}{10}$, and $\frac{2}{10}$, respectively. The probabilities that the bonus scheme will be introduced if they become MD's are 0.02, 0.03, and 0.04. Find the probabilities of A, B, and C to become MD's if bonus scheme is introduced.

(JNTU Aug./Sep. 2008, Nov. 2006)

Solution: Let B_1, B_2, B_3 be the events that A, B, and C become the MD's of a factory.



$P(B_1)$ = Probability that A becomes the MD of the factory

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

$P(B_2)$ = Probability that B becomes the MD of the factory

$$= \frac{3}{10}$$

$P(B_3)$ = Probability that C becomes the MD of the factory

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

Let D be the event that bonus scheme is introduced.

$P\left(\frac{D}{B_1}\right)$ = Probability that bonus scheme will be introduced if B_1 becomes MD of the factory

$$= 0.02$$

$P\left(\frac{D}{B_2}\right)$ = Probability that bonus scheme will be introduced if B_2 becomes M.D of the factory

$$= 0.03$$

$P\left(\frac{D}{B_3}\right)$ = Probability that bonus scheme will be introduced if B_3 becomes MD of the factory

$$= 0.04$$

From theorem of total probability,

$$P(D) = P(B_1) \cdot P\left(\frac{D}{B_1}\right) + P(B_2) \cdot P\left(\frac{D}{B_2}\right) + P(B_3) \cdot P\left(\frac{D}{B_3}\right)$$

$$= \left(\frac{5}{10}\right)(0.02) + \left(\frac{3}{10}\right)(0.03) + \left(\frac{2}{10}\right)(0.04) = 0.027.$$

(i) $P\left(\frac{B_1}{D}\right)$ = Probability of A to become MD's if bonus scheme is introduced

Using Baye's theorem, we have

$$P\left(\frac{B_1}{D}\right) = \frac{P(B_1) \cdot P\left(\frac{D}{B_1}\right)}{P(B_1) \cdot P\left(\frac{D}{B_1}\right) + P(B_2) \cdot P\left(\frac{D}{B_2}\right) + P(B_3) \cdot P\left(\frac{D}{B_3}\right)}$$

$$= \frac{P(B_1) \cdot P\left(\frac{D}{B_1}\right)}{P(D)}$$

$$= \frac{\left(\frac{5}{10}\right)(0.02)}{0.027} = 0.37037$$

(ii) $P\left(\frac{B_2}{D}\right)$ = Probability of B to become MD's if bonus scheme is introduced

Using Baye's theorem, we have

$$P\left(\frac{B_2}{D}\right) = \frac{P(B_2)\cdot P\left(\frac{D}{B_2}\right)}{P(B_1)\cdot P\left(\frac{D}{B_1}\right) + P(B_2)\cdot P\left(\frac{D}{B_2}\right) + P(B_3)\cdot P\left(\frac{D}{B_3}\right)}$$

$$= \frac{P(B_2)\cdot P\left(\frac{D}{B_2}\right)}{P(D)}$$

$$= \frac{\left(\frac{3}{10}\right)(0.03)}{0.027}$$

$$= 0.333333$$

(iii) $P\left(\frac{B_3}{D}\right)$ = Probability of C to become MD if bonus scheme is introduced

Using Baye's theorem, we have

$$P\left(\frac{B_3}{D}\right) = \frac{P(B_3)\cdot P\left(\frac{D}{B_3}\right)}{P(B_1)\cdot P\left(\frac{D}{B_1}\right) + P(B_2)\cdot P\left(\frac{D}{B_2}\right) + P(B_3)\cdot P\left(\frac{D}{B_3}\right)}$$

$$= \frac{P(B_3)\cdot P\left(\frac{D}{B_3}\right)}{P(D)}$$

$$= \frac{\left(\frac{2}{10}\right)(0.04)}{0.027}$$

$$= 0.296296.$$

EXAMPLE 1.13

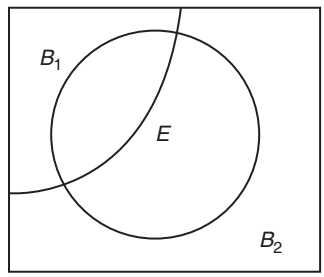
There are two boxes. In box I, 11 cards are there numbered from 1 to 11 and in box II, 5 cards are numbered from 1 to 5. A box is chosen and a card is drawn. If the card shows an even number then another card is drawn from the same box. If card shows an odd number, another card is drawn from the other box. Find the probabilities that

(i) Both are even

(ii) Both are odd

(iii) If both are even, what is the probability that they are from box I.

(JNTU Aug./Sep. 2007, set-2, May 2007, set-3)



Solution: Let B_1 be the event that box I is chosen, B_2 be the event that Box II is chosen.
Let E be the event that card chosen is even and O be the event that card chosen is odd.

(i) Suppose box I is chosen and a card is drawn which is even, and the second card should also be from box I.

From theorem of total probability,

Probability that the card drawn is even = $P(E)$

$$= P(B_1)\cdot P\left(\frac{E}{B_1}\right)+P(B_2)\cdot P\left(\frac{E}{B_2}\right)$$

$P\left(\frac{E}{B_1}\right)$ = Probability of drawing an even card from box I

Since there are 2, 4, 6, 8, 10 = 5 even cards in box I,

$$P\left(\frac{E}{B_1}\right)=\frac{5}{11}$$

$P\left(\frac{E}{B_2}\right)$ = Probability of drawing an even card from box II

Since there are 2, 4 = 2 even cards in box II,

$$P\left(\frac{E}{B_2}\right)=\frac{2}{5}$$

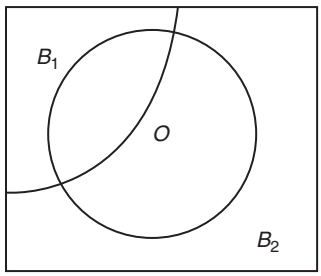
$$P(E)=\left(\frac{1}{2}\right)\cdot\left(\frac{5}{11}\right)+\left(\frac{1}{2}\right)\left(\frac{2}{5}\right)$$

$$=\frac{5}{22}+\frac{1}{5}=\frac{47}{110}.$$

(ii) Suppose box I is chosen and a card is drawn which is odd, then the second card is drawn from box II. From the theorem of total probability,

Probability that the card drawn is odd = $P(O)$

$$= P(B_1)\cdot P\left(\frac{O}{B_1}\right)+P(B_2)\cdot P\left(\frac{O}{B_2}\right)$$



$P\left(\frac{O}{B_1}\right)$ = Probability of drawing an odd card from box I

Since there are 1, 3, 5, 7, 9, 11 = 6 odd cards from box I

$$P\left(\frac{O}{B_1}\right)=\frac{6C_1}{11C_1}=\frac{6}{11}$$

$P\left(\frac{O}{B_2}\right)$ = Probability of drawing an odd card from box II

Since there are 1, 3, 5 = 3 odd cards from box II

$$P\left(\frac{O}{B_1}\right)=\frac{3C_2}{5C_2}=\frac{3}{5}$$

$$P(O) = \left(\frac{1}{2}\right)\cdot\left(\frac{6}{11}\right) + \left(\frac{1}{2}\right)\cdot\left(\frac{3}{5}\right)$$
$$= \frac{63}{110}.$$

(iii) From Baye's theorem

Probability that both even cards are from box I = $P\left(\frac{B_1}{E}\right)$

$$= \frac{P(B_1)\cdot P\left(\frac{E}{B_1}\right)}{P(B_1)\cdot P\left(\frac{E}{B_1}\right) + P(B_2)\cdot P\left(\frac{E}{B_2}\right)}$$

$$P\left(\frac{B_1}{E}\right) = \frac{\left(\frac{1}{2}\right)\cdot\left(\frac{5}{11}\right)}{\frac{47}{110}}$$
$$= \frac{5\times 110}{22\times 47}$$
$$= \frac{25}{47}.$$

EXAMPLE 1.14

Of the three men, the chances that a politician, a businessman, or an academician will be appointed as a vice-chancellor (VC) of a university are 0.5, 0.3, and 0.2, respectively.

(i) Determine the Probability that research is promoted.

(ii) If research is promoted, what is the probability that VC is an academician?

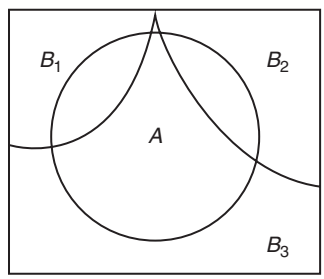
(JNTU Aug./Sep. 2007, set-2, April/May 2007, set-3)

Solution: Let B_1, B_2, and B_3 be the events that politician, businessman, and academician will be appointed as a VC of the university.

$P(B_1)$ = Probability that politician will be appointed as VC of the university
= 0.5

$P(B_2)$ = Probability that businessman will be appointed as VC of the university
= 0.3

$P(B_3)$ = Probability that academician will be appointed as VC of the university
= 0.2



Let A be the event that research is promoted.

$P(A/B_1)$ = Probability that research is promoted if B_1 becomes VC of the university
= 0.3

$P\left(\frac{A}{B_2}\right)$ = Probability that research is promoted if B_2 becomes VC of the university
$= 0.7$

$P\left(\frac{A}{B_3}\right)$ = Probability that research is promoted if B_3 becomes VC of the university
$= 0.8$

From theorem of total probability,

$$P(A) = P(B_1)\cdot P\left(\frac{A}{B_1}\right) + P(B_2)\cdot P\left(\frac{A}{B_2}\right) + P(B_3)\cdot P\left(\frac{A}{B_3}\right)$$

$$= (0.5)(0.3) + (0.3)(0.7) + (0.2)(0.8)$$

$$= 0.52.$$

(i) $P\left(\frac{B_1}{A}\right)$ = Probability of B_1 to become VC if research is promoted

Using Baye's theorem, we have

$$P\left(\frac{B_1}{A}\right) = \frac{P(B_1)\cdot P\left(\frac{A}{B_1}\right)}{P(B_1)\cdot P\left(\frac{A}{B_1}\right) + P(B_2)\cdot P\left(\frac{A}{B_2}\right) + P(B_3)\cdot P\left(\frac{A}{B_3}\right)}$$

$$= \frac{P(B_1)\cdot P\left(\frac{A}{B_1}\right)}{P(D)}$$

$$= \frac{(0.5)(0.3)}{0.52}$$

$$= 0.288462.$$

(ii) $P\left(\frac{B_2}{A}\right)$ = Probability of B_2 to become VC, if research is promoted

Using Baye's theorem, we have

$$P\left(\frac{B_2}{A}\right) = \frac{P(B_2)\cdot P\left(\frac{A}{B_2}\right)}{P(B_1)\cdot P\left(\frac{A}{B_1}\right) + P(B_2)\cdot P\left(\frac{A}{B_2}\right) + P(B_3)\cdot P\left(\frac{A}{B_3}\right)}$$

$$P\left(\frac{B_2}{A}\right) = \frac{P(B_2)\cdot P\left(\frac{A}{B_2}\right)}{P(D)}$$

$$= \frac{(0.3)(0.7)}{0.52}$$

$$= 0.403846.$$

(iii) $P\left(\frac{B_3}{A}\right)$ = Probability of B_3 to become VC, if research is promoted

Using Baye's theorem, we have

$$P\left(\frac{B_3}{A}\right) = \frac{P(B_3)\cdot P\left(\frac{A}{B_3}\right)}{P(B_1)\cdot P\left(\frac{A}{B_1}\right) + P(B_2)\cdot P\left(\frac{A}{B_2}\right) + P(B_3)\cdot P\left(\frac{A}{B_3}\right)}$$

$$= \frac{P(B_3)\cdot P\left(\frac{A}{B_3}\right)}{P(D)}$$

$$= \frac{(0.2)(0.8)}{0.52}$$

$$P\left(\frac{B_3}{A}\right) = 0.307692.$$

EXAMPLE 1.15

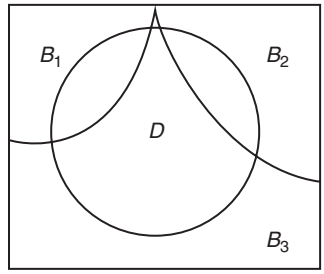
Urn A contains 3 white, 1 black, and 4 red balls. Urn B contains 4 white, 3 black, and 4 red balls. Urn C contains 4 white, 3 black, and 2 red balls. One urn is chosen at random and 2 balls are drawn. They happen to be red and black. What is the probability that both the balls come from urn B?

Solution: Let B_1, B_2, and B_3 be the probabilities of choosing urns A, B, and C.

Let D be the event of drawing one red and one black ball.

Probability of choosing an urn is

$$P(B_1) = P(B_2) = P(B_3) = \frac{1}{3}$$



$P\left(\frac{D}{B_1}\right)$ = Probability of drawing one red and one black ball from urn A

Since urn A contains 3 white, 1 black, and 4 red balls

Total number of balls = (3 + 1 + 4) = 8 balls

$$P\left(\frac{D}{B_1}\right) = \frac{\text{Favourable number of cases}}{\text{Total number of cases}}$$

$$= \frac{\text{Drawing 1 red from 4 red and 1 black from 1 black}}{\text{Drawing two balls from 8 balls}}$$

$$P\left(\frac{D}{B_1}\right) = \frac{4C_1 \cdot 1}{8C_2}$$

$$= \frac{4}{28} = \frac{1}{7}$$

$P\left(\frac{D}{B_2}\right)$ = Probability of drawing one red and one black ball from urn B

Since urn B contains 4 white, 3 black, and 4 red balls,

Total number of balls = (4 + 3 + 4) = 11 balls

$$P\left(\frac{D}{B_2}\right) = \frac{\text{Favourable number of cases}}{\text{Total number of cases}}$$
$$= \frac{\text{Drawing 1 red from 4 red and 1 black from 3 black}}{\text{Drawing two balls from 8 balls}}$$
$$= \frac{4C_1 \cdot 3C_1}{8C_2}$$
$$= \frac{4 \cdot 3}{28} = \frac{12}{28} = \frac{3}{7}$$

$P\left(\frac{D}{B_3}\right)$ = Probability of drawing one red and one black ball from urn C

Since urn C contains 4 white, 3 black, and 2 red balls,

Total number of balls = (4 + 3 + 2) = 9 balls

$$P\left(\frac{D}{B_3}\right) = \frac{\text{Favourable number of cases}}{\text{Total number of cases}}$$
$$= \frac{\text{Drawing 1 red from 2 red and 1 black from 3 black}}{\text{Drawing 2 balls from 8 balls}}$$
$$= \frac{4C_2 \cdot 3C_1}{8C_2}$$
$$= \frac{6 \cdot 3}{28} = \frac{18}{28} = \frac{9}{14}$$

From theorem of total probability,

$$P(D) = P(B_1) \cdot P\left(\frac{D}{B_1}\right) + P(B_2) \cdot P\left(\frac{D}{B_2}\right) + P(B_3) \cdot P\left(\frac{D}{B_3}\right)$$
$$= \left(\frac{1}{3}\right)\left(\frac{1}{7}\right) + \left(\frac{1}{3}\right)\left(\frac{12}{28}\right) + \left(\frac{1}{3}\right)\left(\frac{18}{28}\right)$$
$$P(D) = 0.404762.$$

Probability of both the balls (one red and one black) from urn $B = P\left(\frac{B_2}{D}\right)$

Using Baye's theorem,

$$P\left(\frac{B_2}{D}\right) = \frac{P(B_2) \cdot P\left(\frac{D}{B_2}\right)}{P(B_1) \cdot P\left(\frac{D}{B_1}\right) + P(B_2) \cdot P\left(\frac{D}{B_2}\right) + P(B_3) \cdot P\left(\frac{D}{B_3}\right)}$$
$$= \frac{P(B_2) \cdot P\left(\frac{D}{B_2}\right)}{P(D)}$$
$$= \frac{\left(\frac{1}{3}\right)\left(\frac{12}{18}\right)}{0.404762}$$
$$= 0.352941.$$

EXAMPLE 1.16

There are 3 urns having the following compositions of black and white balls.

Urn I: 6 white, 4 black balls; urn II: 3 white, 7 black balls; urn III: 2 white, 8 black balls.

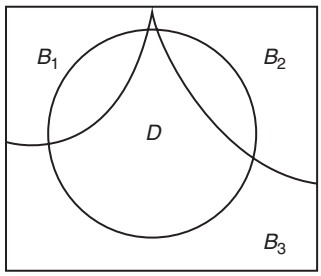
One of these urns is chosen at random with probabilities 0.30, 0.50, and 0.20, respectively. From the chosen urn, two balls are drawn without replacement. Calculate the probabilities that both balls are white.

Solution: Let B_1, B_2, and B_3 be the probabilities of choosing urns I, II, and III.

Probability of choosing urn I = $P(B_1) = 0.30$

Probability of choosing urn II = $P(B_2) = 0.50$

Probability of choosing urn III = $P(B_3) = 0.20$



Let D be the event of drawing two white balls.

$P\left(\frac{D}{B_1}\right)$ = Probability of drawing white balls from urn I

Since urn I contains 6 white and 4 black balls

Total number of balls = (6 + 4) = 10 balls

$$P\left(\frac{D}{B_1}\right) = \frac{\text{Favourable number of cases}}{\text{Total number of cases}}$$

$$= \frac{\text{Drawing 2 white from 6 white balls}}{\text{Drawing two balls from 10 balls}}$$

$$= \frac{6C_2}{10C_2}$$

$$= \frac{6}{10} = \frac{3}{5}$$

$P\left(\frac{D}{B_2}\right)$ = Probability of drawing white balls from urn II

Since urn II contains 3 white and 7 black balls

Total number of balls = (3 + 7) = 10 balls

$$P\left(\frac{D}{B_2}\right) = \frac{\text{Favourable number of cases}}{\text{Total number of cases}}$$

$$= \frac{\text{Drawing 2 white from 3 white balls}}{\text{Drawing 2 balls from 10 balls}}$$

$$= \frac{3C_2}{10C_2}$$

$$= \frac{3}{10}$$

$P\left(\frac{D}{B_3}\right)$ = Probability of drawing white balls from urn III

Since urn III contains 2 white and 8 black balls

Total number of balls = (2 + 8) = 10 balls

$$P\left(\frac{D}{B_3}\right) = \frac{\text{Favourable number of cases}}{\text{Total number of cases}}$$

$$= \frac{\text{Drawing 2 white from 2 white balls}}{\text{Drawing 2 balls from 10 balls}}$$

$$= \frac{2C_2}{10C_2}$$
$$= \frac{1}{45}$$

From theorem of total probability,

Probability that both the balls are white = $P(D)$

$$P(D) = P(B_1)\cdot P\left(\frac{D}{B_1}\right) + P(B_2)\cdot P\left(\frac{D}{B_2}\right) + P(B_3)\cdot P\left(\frac{D}{B_3}\right)$$
$$= (0.3)\left(\frac{3}{5}\right) + (0.5)\left(\frac{3}{10}\right) + (0.2)\left(\frac{1}{45}\right)$$
$$= 0.334444.$$

Work Book Exercises

8. Suppose 5 men out of 100 and 25 women out of 10,000 are colour blind. A colour blind person is chosen at random. What is the probability of the person being a male. (Assume that the number of males and females are equal).

(JNTU April/May 2004, set-3)

9. A business man goes to hotels *X*, *Y*, and *Z* 20%, 50%, and 30% of the times, respectively. It is known that 5%, 4%, and 8% of the rooms in *X*, *Y*, and *Z* hotels have faulty plumbing. What is the probability that businessman's room having faulty plumbing is assigned to hotel *Z*?

(JNTU Aug./Sep. 2005)

10. Three urns are given each containing red and white chips as indicated:

Urn I: 7 red and 3 white; urn II: 3 red and 5 white; urn III: 2 red and 8 white

(i) An urn is chosen at random and a chip is drawn from this urn. The chip is red. Find the probability that the urn chosen was urn I.

(ii) An urn is chosen at random and two balls are drawn without replacement from this urn. If both the balls are red find the probability that urn I was chosen. Under these conditions, what is the probability that urn III was chosen.

11. The chances of *A*, *B*, and *C* in becoming the managers of a company is 4:2:3. The probabilities that their salaries will be revised if *A*, *B*, and *C* become managers are 0.3, 0.5, and 0.8, respectively. If the salaries are revised, what are the probabilities that *A*, *B*, and *C* are appointed as managers?

12. A city is partitioned into districts *A*, *B*, and *C* having 20%, 40%, and 40% of the registered voters, respectively. The registered voters listed as Democrats are 50% in *A*, 25% in *B*, and 75% in *C*.

If a registered voter is chosen randomly in the city, find the probability that he is from

(i) district *A*

(ii) district *B*

(iii) district *C*

13. In a certain college 25% of boys and 10% of girls are studying mathematics. The girls constitute 60% of the students. If a student is selected at random and is found to be studying mathematics, find the probability that the student is a

(i) Girl

(ii) Boy

(JNTU Nov. 2004, set-2)

14. Companies C_1, C_2, and C_3 produce 40%, 35%, and 25% of the cars, respectively. It is known that 3%, 2%, and 2% of the cars produced from C_1, C_2, and C_3 are defective.
    (i) What is the probability that a car purchased is found to be defective?
    (ii) If a car purchased to be defective what is the probability that this car is produced by company C_3?
15. A class contains 10 boys and 5 girls. Three students are selected at random one after another. Find the probability that
    (i) First two are boys and third is a girl.
    (ii) First and third are of the same sex and second is of opposite sex.

(JNTU 2004S, Feb. 2008S)

DEFINITIONS AT A GLANCE

Random Experiment: An experiment whose outcome or results are not unique and which cannot be predicted with certainty is called random experiment.

Sample Space: The set of all outcomes of a random experiment is called sample space and is denoted by *S*.

Mutually Exclusive Events: Two events *A* and *B* are said to be mutually exclusive events if the occurrence of one event excludes (precludes) the occurrence of the other event.

Collectively Exhaustive Events: A list of events A_1, A_2, A_3, A_4,… A_n are collectively exhaustive if the union of all the events is the entire sample space.

Equally Likely Events: The events *A* and *B* are said to be equally likely events if each of the elements have equal chance of occurrence.

Mathematical or Classical Probability: If an event *E* can occur in *m* ways out of *n* mutually exclusive, equally likely, and collectively exhaustive ways, then the probability of occurrence of event *E*, denoted by *P*(*E*) is given by

$$P(E) = \frac{\text{Favourable number of cases for } E}{\text{Total number of cases}}$$

$$P(E) = \frac{m}{n}$$

Conditional Probability: The probability of an event occurring under a condition is called conditional probability. If an event *A* has already occurred, then the probability of another event *B* with the condition that *A* has already occurred is called the conditional probability of *B* given *A*, denoted by $P\left(\frac{B}{A}\right)$.

Independent Events: Two events are said to be independent if the occurrence of one event affects the occurrence of the other event.

FORMULAE AT A GLANCE

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

- $nC_r = \frac{nP_r}{r!}$
- If $A \subseteq B$, then $P(A) \le P(B)$
- If $A \cap B = \emptyset$, then $P(A) \le P(B^c)$
- If A is any event of S then $0 \le P(E) \le 1$
- If A is any event in S, then $P(A^c) = 1 - P(A)$
- For any two events A and B, $P(A^c \cap B) = P(B) - P(A \cap B)$
- **Addition Theorem:** For any two events, A and B of S, $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- For any three events A, B, and C

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

- The conditional probability of B given A is $P\left(\frac{B}{A}\right) = \frac{P(A \cap B)}{P(A)}$
- **Multiplication Theorem:** For any two events A and B, $P(A \cap B) = P(A)$. $P\left(\frac{B}{A}\right)$, $P(A) > 0$.
- **Independent Events:** If two events A and B are independent events, then $P(A \cap B) = P(A) \cdot P(B)$.
- **Theorem of Total Probability:** Let $B_1, B_2, B_3, \ldots B_k$ constitute a partition of the sample space S with $P(B_i) \neq 0$ for any $i = 1, 2, \ldots k$. Then for any event A of S,

$$P(A) = \sum_{i=1}^{k} P(B_i \cap A)$$
$$= \sum_{i=1}^{k} P(B_i) P(A/B_i)$$

- **BAYE'S Theorem:** Let the Sample space S be partitioned into k subsets $B_1, B_2, \ldots B_k$ with $P(B_i) \neq 0$. For $i = 1, 2, \ldots k$. For any arbitrary event A in S, with $P(A) \neq 0$ and any specific event B_s, we have

$$P\left(\frac{B_s}{A}\right) = \frac{P(B_s \cap A)}{\sum_{i=1}^{n} P(B_i \cap A)}$$
$$= \frac{P(B_s) \cdot P\left(\frac{A}{B_s}\right)}{\sum_{i=1}^{k} P(B_i) \cdot P\left(\frac{A}{B_i}\right)} \quad \text{for } s = 1, 2, \ldots k$$

OBJECTIVE TYPE QUESTIONS

1. The odds in favour of drawing a king or a diamond from a well shuffled pack of 52 cards are
   (a) 9:4 (b) 4:9
   (c) 5:9 (d) 9:5
2. The probability of getting a number greater than 2 or an even number in a single of a fair die is
   (a) $\frac{2}{3}$ (b) $\frac{1}{3}$
   (c) $\frac{5}{6}$ (d) none

3. If there are three mutually exclusive events A, B, and C with $P(A) = 2\ P(B) = 3\ P(C)$, then $P(A) =$ _____.

   (a) $\frac{2}{11}$ (b) $\frac{3}{11}$

   (c) $\frac{6}{11}$ (d) $\frac{5}{11}$

4. If A and B are independent events such that $P(A) = 0.2$, $P(A \cup B) = 0.8$, then $P(B) =$ _____.

   (a) $\frac{3}{4}$ (b) $\frac{1}{2}$

   (c) $\frac{1}{4}$ (d) 1

5. If a card is drawn from a well shuffled pack of 52 cards, then the probability that it is a spade or a queen is _____.

   (a) $\frac{2}{13}$ (b) $\frac{5}{13}$

   (c) $\frac{3}{13}$ (d) $\frac{4}{13}$

6. Six boys and six girls sit around a table randomly. The probability that all the six girls sit together is _____.

   (a) $\frac{1}{77}$ (b) $\frac{2}{77}$

   (c) $\frac{3}{77}$ (d) $\frac{4}{77}$

7. If for any two events A and B the following are correct: $P(A) = P\left(\frac{A}{B}\right) = \frac{1}{4}$ and $\frac{1}{2}$, then

   (a) A and B are mutually exclusive (b) A and B are independent

   (c) $P\left(\frac{A^c}{B}\right) = \frac{3}{4}$ (d) none

8. A box contains n tickets marked 1 through n. Two tickets are drawn in succession without replacement. The probability that the numbers on the tickets are consecutive integers is _____.

   (a) $\frac{1}{(n-1)}$ (b) $\frac{1}{n}$

   (c) $\frac{1}{(n+1)}$ (d) none

9. The probability for a leap year to have 52 Mondays and 53 Sundays

   (a) $\frac{1}{7}$ (b) $\frac{1}{5}$

   (c) $\frac{1}{3}$ (d) none

10. An urn contains 2 red balls, 6 white balls, and 6 black balls, then the probability of drawing a red or a black ball is _____.

(a) $\frac{3}{7}$ (b) $\frac{4}{7}$

(c) $\frac{5}{7}$ (d) $\frac{6}{7}$

ANSWERS

1. (b) 2. (c) 3. (c) 4. (a) 5. (a) 6. (a) 7. (b) 8. (a)
9. (a) 10. (b)

2 Random Variables (Discrete and Continuous)

Prerequisites

Before you start reading this unit, you should:

- Have some knowledge on definite integrals
- Know to solve double integrals (multiple integrals)
- Know about probability and calculating it for simple problems

Learning Objectives

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

- Know and differentiate between discrete and continuous random variables
- Know the density and distribution functions for discrete and continuous variables
- Know about joint probability distributions
- Know about stochastic independence of random variables
- Have an idea of transformation of one and two dimensional random variables

INTRODUCTION

In the previous Unit, we are familiar with some new terms like random experiment, events, sample space, etc. Now, we shall know about a random variable, the types of a random variable, and the concepts attached to it. By a random variable, we mean a real number associated with the outcome of a random experiment.

Example: Suppose two coins are tossed simultaneously, then the sample space will be $S = \{HH, HT, TH, TT\}$. Let a variable X denotes the number of heads. The probability distribution of this variable is as follows:

$X = x$	0	1	2
$P(X = x)$	$\frac{1}{4}$	$\frac{2}{4}$	$\frac{1}{4}$

2.1 RANDOM VARIABLE

A random variable X on a sample space S is a rule that assigns a numerical value to each element of S, that is, a random variable is a function from the sample space S into the set of real numbers R.

These are of the following two types:

1. **Discrete random variables:** A random variable which assumes integral values only in an interval of domain is called discrete random variable.

   Example: (i) The number of rooms in the houses of a township.

   (ii) The number of children in the families of a colony.

2. **Continuous random variables:** Quantities which are capable of taking all possible values in a certain range are called continuous random variables.
   Example: (i) Weights of students in a class.
   (ii) The height of a child as he grows.

   After knowing about discrete and continuous random variables, we shall know the functions and distributions attached to them.

2.2 PROBABILITY MASS FUNCTION (PMF)

Let X be a discrete random variable taking values $x = 0, 1, 2, \ldots$ then $P(X = x)$ is called pmf, if it satisfies the following properties:

(i) $P(X = x) \geq 0$

(ii) $\sum_{x=0}^{\infty} P(X = x) = 1.$

2.3 PROBABILITY DENSITY FUNCTION (PDF)

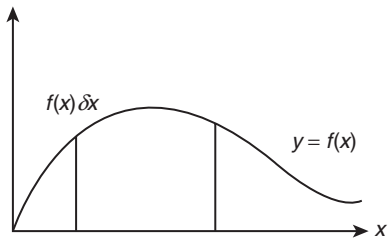
Let X be a continuous random variable taking values in a certain range $a \leq X \leq b$ then the function $P(X = x) = f(x)$ is called pdf, if it satisfies the following properties:

(i) $f(x) \geq 0$

(ii) $\int_a^b f(x)\,dx = 1.$

Caution:

- Total area under the probability curve $y = f(x)$ is unity.
- $f(x) \geq 0$ implies that the graph of $f(x)$ is above the x-axis.



Area under the probability curve

Worked Out Examples

EXAMPLE 2.1

For the discrete probability distribution

x	0	1	2	3	4	5	6	7
$f(x)$	0	K	$2K$	$2K$	$3K$	K^2	$2K^2$	$7K^2 + K$

Determine the following:

(i) The value of K

(ii) Probability that $X < 6$

(iii) $P(X \geq 6)$

(iv) $P(0 < X < 5)$

(v) Distribution function of X.

Solution:

(i) Since given $f(x)$ is a probability mass function $\sum f(x) = 1$

$\Rightarrow 0 + K + 2K + 2K + 3K + K^2 + 2K^2 + 7K^2 + K = 1$

$10K^2 + 9K - 1 = 1,$

$$K = -1, \frac{1}{10}$$

$\because$ K cannot be negative, $K = \frac{1}{10}$

Hence, with this value of K, the probability distribution function becomes

x	0	1	2	3	4	5	6	7
$P(X = x)$	0	$\frac{1}{10}$	$\frac{2}{10}$	$\frac{2}{10}$	$\frac{3}{10}$	$\frac{1}{100}$	$\frac{2}{100}$	$\frac{17}{100}$

(ii) $P(X < 6) = P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4) + P(X = 5)$

$$= 0 + \frac{1}{10} + \frac{2}{10} + \frac{2}{10} + \frac{3}{10} + \frac{1}{100} + \frac{2}{100}$$

$$= \frac{81}{100}$$

(iii) $P(X \geq 6) = 1 - P(X < 6)$

$$= 1 - \frac{81}{100}$$

$$= \frac{19}{100}$$

(iv) $P(0 < X < 5) = P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4)$

$$= \frac{1}{10} + \frac{2}{10} + \frac{2}{10} + \frac{3}{10}$$

$$= \frac{8}{10}$$

$$= \frac{4}{5}$$

(v) The distribution function of X is

X	0	1	2	3	4	5	6	7
$P(X = x)$	0	$\frac{1}{10}$	$\frac{2}{10}$	$\frac{2}{10}$	$\frac{3}{10}$	$\frac{1}{100}$	$\frac{2}{100}$	$\frac{17}{100}$
$F(x) = P(X \leq x)$	0	$\frac{1}{10}$	$\frac{3}{10}$	$\frac{5}{10}$	$\frac{8}{10}$	$\frac{81}{100}$	$\frac{83}{100}$	1

EXAMPLE 2.2

If the probability density of a random variable is given by

$$f(x) = \begin{cases} K(1-x^2), \text{ for } 0 < X < 1. \\ 0, \qquad \text{otherwise} \end{cases}$$

Find the value of K and the probabilities that a random variable will take on a value

(i) Between 0.1 and 0.2
(ii) Greater than 0.5
(iii) Mean
(iv) Variance

Solution:

(i) Since $f(x)$ is a pdf we have

$$\int_{-\infty}^{\infty} f(x)\,dx = 1$$

$$\int_{-\infty}^{0} f(x)\,dx + \int_{0}^{1} f(x)\,dx + \int_{1}^{\infty} f(x)\,dx = 1$$

$$\therefore \int_{0}^{1} f(x)\,dx = 1$$

$$K\int_{0}^{1} (1-x^2)\,dx = 1$$

$$K\left[x - \frac{x^3}{3}\right]_{x=0}^{1} = 1; K\left[1 - \frac{1}{3}\right] = 1;$$

$$K\left(\frac{2}{3}\right) = 1$$

$$\therefore K = \frac{3}{2}, \therefore f(x) = \frac{3}{2}(1-x^2)$$

(ii) Probability that X will take on a value between 0.1 and 0.2 is

$$= P(0.1 < X < 0.2)$$

$$P(0.1 < X < 0.2) = \int_{0.1}^{0.2} f(x)\,dx$$

$$= \frac{3}{2}\int_{0.1}^{0.2} (1-x^2)\,dx = \frac{3}{2}\left|x - \frac{x^3}{3}\right|_{0.1}^{0.2}$$

$$P(0.1 < X < 0.2) = \frac{3}{2}\left[(0.2 - 0.1) - \frac{1}{3}(0.2^3 - 0.1^3)\right]$$

$$= \frac{3}{2}[0.1 - 0.0023]$$

$$P(0.1 < X < 0.2) = 0.1465$$

(iii) Probability that X will take a value greater than 0.5

$$= P(X > 0.5)$$

$$= \int_{0.5}^{1} f(x)\,dx = \int_{0.5}^{1} \frac{3}{2}(1-x^2)\,dx$$

$$= \frac{3}{2}\left| x - \frac{x^3}{3} \right|_{0.5}^{1}$$

$$= \frac{3}{2}\left[(1-0.5) - \frac{0.5^3}{3} \right]$$

$$P(X > 0.5) = \frac{3}{2}[0.5 - 0.0416] = 0.687$$

(iv) Mean of $X = E(X) = \int_{-\infty}^{\infty} x f(x)\,dx$

$$= \frac{3}{2}\int_0^1 x(1-x^2)\,dx = \frac{3}{2}\int_0^1 (x - x^3)\,dx$$

$$= \frac{3}{2}\left[\frac{x^2}{2} - \frac{x^4}{4} \right]_0^1 = \frac{3}{2}\left[\frac{1}{2}(1-0) - \frac{1}{4}(1-0) \right]$$

$$E(X) = \frac{3}{2}\left[\frac{1}{2} - \frac{1}{4} \right] = \frac{3}{2} \times \frac{1}{4} = \frac{3}{8}$$

$$E(X^2) = \int_{-\infty}^{\infty} x^2 f(x)\,dx = \frac{3}{2}\int_0^1 x^2(1-x^2)\,dx$$

$$= \frac{3}{2}\int_0^1 (x^2 - x^4)\,dx$$

$$= \frac{3}{2}\left[\frac{x^3}{3} - \frac{x^5}{5} \right]_0^1$$

$$E(X^2) = \frac{3}{2}\left[\frac{1}{3} - \frac{1}{5} \right] = \frac{3}{2}\left[\frac{5-3}{15} \right] = \frac{3}{2} \times \frac{2}{15}$$

$$\therefore E(X^2) = \frac{3}{15} = \frac{1}{5}$$

Variance of $X = V(X) = E(X^2) - [E(X)]^2$

$$= \frac{1}{5} - \left(\frac{3}{8} \right)^2 = 0.059$$

EXAMPLE 2.3

On a laboratory assignment, if the equipment is working, the density function of the observed outcome, X is

$$f(x) = \begin{cases} 2(1-x), & 0 < x < 1 \\ 0, & \text{otherwise} \end{cases}$$

(i) Calculate $P\left(X \le \frac{1}{3}\right)$.

(ii) What is the probability that X will exceed 0.5?

(iii) Given that $X \ge 0.5$, what is the probability that X will be less than 0.75?

Solution:

(i) $$P\left(X \le \frac{1}{3}\right) = \int_0^{1/3} f(x)\,dx = 2\int_0^{1/3} (1-x)\,dx$$

$$= 2\left|x - \frac{x^2}{2}\right|_0^{1/3} = 2\left[\frac{1}{3} - \frac{1}{2}\left(\frac{1}{3}\right)^2\right]$$

$$P\left(X \le \frac{1}{3}\right) = 0.1388$$

(ii) Probability that X will exceed $0.5 = P(X > 0.5)$

$$= \int_{0.5}^{1} f(x)\,dx = 2\int_{0.5}^{1} (1-x)\,dx$$

$$= 2\left|x - \frac{x^2}{2}\right|_0^{1} = 2\left[(1-0.5) - \frac{1}{2}(1-0.5^2)\right]$$

$$P(X > 0.5) = 2[0.5 - 0.375] = 0.25$$

(iii) Given $X \ge 0.5$, probability that X will be less than 0.75

$$= P(X < 0.75) = \int_{0.5}^{0.75} f(x)\,dx = 2\int_{0.5}^{0.75} (1-x)\,dx$$

$$= 2\left|x - \frac{x^2}{2}\right|_{0.5}^{0.75} = 2\left[(0.75-0.5) - \frac{1}{2}(0.75^2 - 0.5^2)\right]$$

$$= 2[0.25 - 0.15625] = 0.1875$$

EXAMPLE 2.4

The amount of bread (in hundreds of pounds) x that a certain bakery is able to sell in a day is found to be a numerical valued random phenomenon with probability density $f(x)$ given by

$$f(x) = \begin{cases} Kx, & 0 \le x < 5 \\ K(10-x), & 5 \le x < 10 \\ 0, & \text{elsewhere} \end{cases}$$

Find K and hence the probability that number of pounds of bread that will be sold next day is

(i) More than 500 pounds

(ii) Less than 500 pounds

(iii) Between 250 and 750 pounds

Solution: Since $f(x)$ is a probability density function,

$$\int_{-\infty}^{\infty} f(x)\,dx = 1 \Rightarrow \int_{0}^{10} f(x)\,dx = 1$$

$$\int_{0}^{5} f(x)\,dx + \int_{5}^{10} f(x)\,dx = 1$$

$$\Rightarrow K\int_{0}^{5} x\,dx + K\int_{5}^{10} (10-x)\,dx = 1$$

$$\Rightarrow K\frac{x^2}{2}\Big|_{0}^{5} + K\left|10x - \frac{x^2}{2}\right|_{5}^{10} = 1$$

$$\Rightarrow K\left[\frac{25}{2} + 10(10-5) - \frac{1}{2}(100-25)\right] = 1$$

$$K\left[\frac{25}{2} + 50 + \frac{75}{2}\right] = 1 \Rightarrow K = \frac{1}{25}$$

(i) Probability that bread sold is more than 500 pounds

$$= P(X > 5) = \int_{5}^{10} f(x)\,dx$$

$$= \frac{1}{25}\int_{5}^{10} (10-x)\,dx = \frac{1}{25}\left[10x - \frac{x^2}{2}\right]_{5}^{10}$$

$$= \frac{1}{25}\left[10(10-5) - \frac{1}{2}(100-25)\right]$$

$$= \frac{1}{25}\left[50 - \frac{75}{2}\right] = \frac{1}{2}$$

(ii) Probability that bread sold is less than 500 pounds = $P(X < 5)$

$P(X \le 5) = 1 - P(X > 5) = 1 - \frac{1}{2} = \frac{1}{2}$ {from the previous (i)}

(iii) Probability that bread sold is between 250 and 750 pounds

$$= P(250 < X < 750) = P(2.5 < X < 7.5) = \int_{2.5}^{7.5} f(x)\,dx$$

$$= \int_{2.5}^{5} f(x)\,dx + \int_{5}^{7.5} f(x)\,dx = \frac{1}{25}\int_{2.5}^{5} x\,dx + \frac{1}{25}\int_{5}^{7.5} (10-x)\,dx$$

$$= \frac{1}{25}\left|\frac{x^2}{2}\right|_{2.5}^{5} + \frac{1}{25}\left|10x - \frac{x^2}{2}\right|_{5}^{7.5} = \frac{1}{25}\left(\frac{25}{2} - \frac{6.25}{2}\right) + \frac{1}{25}[(75-50)-(43.75)]$$

$$\Rightarrow P(250 < X < 750) = 0.75.$$

EXAMPLE 2.5

Let the phase error in a tracking device have probability density $F(x) = \begin{cases} \cos x, & 0 < x < \frac{\pi}{2} \\ 0, & \text{elsewhere} \end{cases}$
Find the probability that the phase error is

(i) Between 0 and $\frac{\pi}{4}$

(ii) Greater than $\frac{\pi}{3}$.

Solution: Let X denotes the phase error in a tracking device.

(i) Probability that phase error is between 0 and $\frac{\pi}{4}$

$$= P\left(0 < X < \frac{\pi}{4}\right) = \int_{0}^{\frac{\pi}{4}} f(x)\,dx = \int_{0}^{\frac{\pi}{4}} \cos x\,dx$$

$$= \sin x\Big|_{0}^{\frac{\pi}{4}} = \frac{1}{\sqrt{2}} = 0.707$$

(ii) Probability that phase error is greater than $\frac{\pi}{3}$.

$$= P\left(X > \frac{\pi}{3}\right) = \int_{\frac{\pi}{3}}^{\infty} \cos x\,dx = \int_{\frac{\pi}{3}}^{\frac{\pi}{2}} \cos x\,dx$$

$$= \sin x\Big|_{\frac{\pi}{3}}^{\frac{\pi}{2}} = 1 - \frac{\sqrt{3}}{2} = 0.133$$

EXAMPLE 2.6

Given $f(x) = \begin{cases} 0, & x < -a \\ \frac{1}{2}\left(\frac{x}{a} + 1\right), & -a \le x \le a \\ 1, & x > a \end{cases}$

(i) Verify if $F(x)$ is a distribution function

(ii) Find the probability density function

(iii) Find probability that X lies between $\frac{-a}{2}$ and $\frac{a}{2}$

(iv) Find also probability that X is greater than $\frac{-a}{2}$

Solution:

(i) Since $F(x) = 1$ for $x > a$

(i.e.,) $F(\infty) = 1$ and $F(x) = 0$ for $x < -a$ (i.e.,) $F(-\infty) = 0$.

(i.e.,) $f(\infty) = \int_{-\infty}^{\infty} f(x)dx = 1$

In addition, $F(-\infty) = \int_{-\infty}^{-\infty} f(x)dx = 0$

Hence, the given $F(x)$, is $0 \le F(x) \le 1$ is a distribution function.

(ii) The density function is given by $f(x) = \frac{d}{dx}F(x)$

$$f(x) = \frac{d}{dx}\left[\frac{1}{2}\left(\frac{x}{a}+1\right)\right], \quad -a \le x \le a$$

$$f(x) = \begin{cases} \frac{1}{2}\ \frac{1}{a}, & -a \le x \le a \\ 0, & \text{otherwise} \end{cases}$$

(iii) Probability that X lies between $\frac{-a}{2}$ and $\frac{a}{2}$ is

$$P\left(\frac{-a}{2} < X < \frac{a}{2}\right) = \int_{\frac{-a}{2}}^{\frac{a}{2}} f(x)dx = \frac{1}{2a}\int_{\frac{-a}{2}}^{\frac{a}{2}} dx$$

$$= \frac{1}{2a}\left|x\right|_{\frac{-a}{2}}^{\frac{a}{2}} = \frac{1}{2a}\left[\frac{a}{2}+\frac{a}{2}\right] = \frac{1}{2}$$

(iv) Probability that X is greater than $\frac{-a}{2} = P\left(X > \frac{-a}{2}\right)$

$$= \int_{\frac{-a}{2}}^{\infty} f(x)dx = \int_{\frac{-a}{2}}^{\frac{a}{2}} \frac{1}{2a}dx = \frac{1}{2}$$

Caution: (iii) and (iv) are the same probabilities.

EXAMPLE 2.7

The proportion of budgets for a certain type of industrial company that is allotted to environmental and pollution control is coming under scrutiny. A data collection project that determines the distribution of these proportions is given by

$$f(x) = \begin{cases} 5(1-x)^4, & 0 \le x \le 1 \\ 0, & \text{otherwise} \end{cases}$$

(i) Verify that the above is a valid density.

(ii) What is the probability that a company chosen at random expends less than 10% of its budget on environmental and pollution controls?

(iii) What is the probability that a company selected at random spends more than 50% on environmental and pollution control?

Solution:

(i) To verify the given $f(x)$ is a density,

$$\int_{-\infty}^{\infty} f(x)dx = 1$$

$$5\int_{0}^{1}(1-x)^4 dx = 5\ \frac{(1-x)^5}{-5}\Bigg|_0^1 = 1$$

Hence, $f(x)$ is a valid density function.

(ii) Let X denotes the proportion of budgets allotted to environmental and pollution control.
Probability that company chosen expends less than 10% of its budget = $P(X < 0.1)$

$$= \int_0^{0.1} f(x)dx = 5\int_0^{0.1}(1-x)^4 dx$$

$$= 5\ \frac{(1-x)^5}{-5}\Bigg|_0^{0.1} = [-(1-0.1)^5 + 1]$$

$$= 0.4095$$

(iii) Probability that a company selected, spends more than 50% on environmental and pollution control = $P(X > 0.50)$

$$\mathrm{P}(X > 0.50) = \int_{0.5}^{1} f(x)dx = 5\int_{0.5}^{1}(1-x)^4 dx$$

$$= 5\ \frac{(1-x)^5}{-5}\Bigg|_{0.5}^{1} = [-(1-1) + (1-0.5)^5]$$

$$= 0.03125$$

Work Book Exercises

1. On a laboratory assignment, if the equipment is working, the density function of the observed outcome X is $f(x) = \begin{cases} 2(1-x), & 0 < x < 1 \\ 0, & \text{otherwise} \end{cases}$

   (i) Compute $\mathrm{P}\left(X > \frac{1}{3}\right)$.

   (ii) What is the probability that X will exceed 0.5 and less than 0.9?

2. A random variable X has the following probability function:

X	1	2	3	4	5	6
$P(X)$	K	$3K$	$5K$	$7K$	$9K$	$11K$

(i) Find the value of K

(ii) Find $P(X \geq 3)$

(iii) Find $P(1 < X \leq 5)$

3. The length of time (in minutes) that a lady speaks on the telephone is found to be random phenomenon, with a probability function given by

$$f(x) = \begin{cases} A e^{-x/5}, & x \geq 0 \\ 0, & \text{otherwise} \end{cases}$$

(i) Find the value of A such that $f(x)$ is a pdf

(ii) What is the probability that the number of minutes the lady talks over the phone is

(a) More than 10 minutes

(b) Less than 5 minutes

(c) Between 5 and 10 minutes

4. A random variable X has the following probability distribution:

Values of X, x	0	1	2	3	4	5	6	7	8
$P(x)$	K	$3K$	$5K$	$7K$	$9K$	$11K$	$13K$	$15K$	$17K$

(i) Determine the value of K

(ii) Find $P(X < 3)$, $P(X \geq 3)$ $P(0 < X < 5)$

(iii) Find the distribution function of X

5. A random variable X has the probability function:

Values of X, x	–2	–1	0	1	2	3
$P(x)$	0.1	K	0.2	$2K$	0.3	K

(i) Determine the value of K

(ii) Construct cumulative distribution function cdf of X.

2.4 JOINT PROBABILITY DISTRIBUTIONS

So far in the previous sections, we have dealt with one dimensional sample spaces and sometimes we get situations where two dimensional samples spaces are to be used.

For example, we might measure the amount of precipitate P and volume V of a gas released from a controlled chemical experiment, which can be thought of as a two dimensional random variable.

If X and Y are two discrete random variables, then the probability distribution for their simultaneous occurrence is represented by a function $f(x, y)$. This function is called joint probability distribution of X and Y.

Hence when X and Y are discrete,

$$p_{ij} = f(x_i, y_j) = P(X = x_i, Y = y_j)$$
$$= P(X = x_i \cap Y = y_j)$$

X \ Y	y_1	$y_2 \cdots$	$y_j \cdots$	y_n	Total
x_1	p_{11}	$p_{12} \cdots$	$p_{1j} \cdots$	p_{1n}	$p_{1.}$
$x_2 \cdots$	$p_{21} \cdots$	$p_{22} \cdots$	$p_{2j} \cdots$	p_{2n}	$p_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$x_i \cdots$	$p_{i1} \cdots$	$p_{i2} \cdots$	$p_{ij} \cdots$	p_{in}	$p_{i.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_m	p_{m1}	$p_{m2} \cdots$	$p_{mj} \cdots$	p_{mn}	$p_{m.}$
Total	$p_{\bullet 1}$	$p_{\bullet 2} \cdots$	$p_{\bullet j} \cdots$	$p_{\bullet n}$	1

Definition

The function $P(x_i, y_j)$ is a joint probability distribution or probability mass function of the discrete random variables X and Y if

(i) $P(x_i, y_j) \geq 0$ for all (x_i, y_j)

(ii) $\sum_{i=1}^{m} \sum_{j=1}^{n} P(x_i, y_j) = 1$

(iii) $P(X = x_i, Y = y_j) = P(X = x_i \cap Y = y_j) = P(x_i, y_j)$

Marginal Probability Functions

The marginal probability function of X is given by

$$p_i = p_{i1} + p_{i2} + \ldots + p_{ij} + \ldots + p_{in}$$

$$= P[X = x_i \cap Y = y_1] + P[X = x_i \cap Y = y_2] + \ldots + P[X = x_i \cap Y = y_j] + \ldots + P[X = x_i \cap Y = y_n]$$

$$= \sum_{j=1}^{n} p(x_i, y_j)$$

In addition, $\sum_{i=1}^{m} p_{i.} = p_{1.} + p_{2.} + \ldots + p_{m.} = 1$

Similarly, the marginal probability function of Y is given by,

$$p_j = p_{1j} + p_{2j} + \ldots + p_{ij} + \ldots + p_{mj}$$

$$= P[X = x_1 \cap Y = y_j] + P[X = x_2 \cap Y = y_j] + \ldots + P[X = x_i \cap Y = y_j] + \ldots + P[X = x_m \cap Y = y_j]$$

$$= \sum_{i=1}^{m} p(x_i, y_j)$$

In addition, $\sum_{j=1}^{n} p_{.j} = p_{.1} + p_{.2} + \cdots + p_{.n} = 1$

Conditional Probability Functions

The conditional probability function of X given $Y = y_j$ is given by

$$P\left[\frac{X = x_i}{Y = y_j}\right] = \frac{P[X = x_i \cap Y = y_j]}{P[Y = y_j]}$$

$$= \frac{p(x_i, y_j)}{p_{.j}}$$

$$= \frac{p_{ij}}{p_{.j}}$$

Similarly, the conditional probability function of Y given $X = x_i$ is given by

$$P\left[\frac{Y = y_i}{X = x_j}\right] = \frac{P[X = x_i \cap Y = y_j]}{P[X = x_i]}$$

$$= \frac{p(x_i, y_j)}{p_{i.}}$$

$$= \frac{p_{ij}}{p_{i.}}$$

2.5 JOINT DENSITY FUNCTION $f(x, y)$

Definition

When X and Y are continuous random variables, the joint density function $f(x, y)$ satisfies the following:

(i) $f(x, y) \geq 0$ for all (x, y)

(ii) $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\,dx\,dy = 1$

The marginal probability function of X and Y are given $f_Y(y) = \int_{-\infty}^{\infty} f_{XY}(x,y)\,dx$

and $f_X(x) = \int_{-\infty}^{\infty} f(x,y)\,dy$

The marginal distribution functions are given by,

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

$= P(X \leq x,\ Y < \infty) = \sum_y P(X \leq x, Y = y)$

$F_Y(y) = P(Y \leq y)$

$= P(X < \infty,\ Y \leq y)$

$= \sum_x P(X = x, Y \leq y)$ for discrete X, Y random variables.

In case of continuous random variables,

$$F_X(x) = \int_{-\infty}^{x}\left[\int_{-\infty}^{\infty} f_{XY}(x,y)\,dy\right]dx$$

$$F_Y(y) = \int_{-\infty}^{y} \left[\int_{-\infty}^{\infty} f_{XY}(x,y)\,dx \right] dy$$

Let us use the above concepts in the following problems.

Worked Out Examples

EXAMPLE 2.8

Suppose that X and Y have the following joint probability distribution

$f(x, y)$ / Y	2	4
1	0.10	0.15
3	0.20	0.30
5	0.10	0.15

(i) Find the marginal distribution of X.

(ii) Find the marginal distribution of Y.

Solution: The marginal distribution of X is

$$P(X=2) = P(X=2, Y=1) + P(X=2, Y=3) + P(X=2, Y=5)$$
$$= 0.10 + 0.20 + 0.10$$
$$= 0.40$$
$$P(X=4) = P(X=4, Y=1) + P(X=4, Y=3) + P(X=4, Y=5)$$
$$= 0.15 + 0.30 + 0.15$$
$$= 0.60$$

The marginal distribution of X is

X	2	4	Total
$f_x(x)$	0.4	0.6	1

The marginal distribution of Y is

$$P(Y=1) = P(X=2, Y=1) + P(X=4, Y=1)$$
$$= 0.10 + 0.15$$
$$= 0.25$$
$$P(Y=3) = P(X=2, Y=3) + P(X=4, Y=3)$$
$$= 0.20 + 0.30$$
$$= 0.50$$
$$P(Y=5) = P(X=2, Y=5) + P(X=4, Y=5)$$
$$= 0.10 + 0.15$$
$$= 0.25$$

The marginal distribution of Y is

Y	1	3	5	Total
$F_y(Y)$	0.25	0.50	0.25	1.0

EXAMPLE 2.9

Let X denote the number of times a certain numerical control machine will malfunction: 1, 2, or 3 times on any given day. Let Y denotes the number of times a technician is called on an emergency call. Their joint probability distribution is given by

$f(x, y)$ $x \rightarrow$ / $y \downarrow$	1	2	3
1	0.05	0.05	0.1
2	0.05	0.1	0.35
3	0	0.2	0.1

(i) Find the marginal distribution of X.

(ii) Find the marginal distribution of Y.

(iii) Find $P\left(\frac{Y=3}{X=2}\right)$.

Solution:

(i) The marginal distribution of X is

$$P(X=1) = P(X=1, Y=1) + P(X=1, Y=2) + P(X=1, Y=3)$$
$$= 0.05 + 0.05 + 0$$
$$= 0.1$$
$$P(X=2) = P(X=2, Y=1) + P(X=2, Y=2) + P(X=2, Y=3)$$
$$= 0.05 + 0.1 + 0.2$$
$$= 0.35$$
$$P(X=3) = P(X=3, Y=1) + P(X=3, Y=2) + P(X=3, Y=3)$$
$$= 0.1 + 0.35 + 0.1$$
$$= 0.55$$

The distribution of X is given by

X	1	2	3	Total
$f_x(x)$	0.1	0.35	0.55	1

(ii) The marginal distribution of Y is

$$P(Y=1) = P(X=1, Y=1) + P(X=2, Y=1) + P(X=3, Y=1)$$
$$= 0.05 + 0.05 + 0.1$$
$$= 0.2$$
$$P(Y=2) = P(X=1, Y=2) + P(X=2, Y=2) + P(X=3, Y=2)$$
$$= 0.05 + 0.1 + 0.35$$
$$= 0.5$$
$$P(Y=3) = P(X=1, Y=3) + P(X=2, Y=3) + P(X=3, Y=3)$$
$$= 0 + 0.2 + 0.1$$
$$= 0.3$$

The distribution of Y is

Y	1	2	3	Total
$f_y(y)$	0.2	0.5	0.3	1

(iii) The conditional distribution of $Y = 3$ given $X = 2$ is

$$P\left(\frac{Y=3}{X=2}\right) = \frac{P(X=2, Y=3)}{P(X=2)}$$

$$\frac{0.2}{0.35} = \frac{2}{35} = 0.5714$$

EXAMPLE 2.10

X and Y are two random values having the joint density function $f(x, y) = \frac{1}{27}(2x + y)$, where X and Y assume integer values 0, 1, and 2. Find the conditional distribution of Y for $X = x$.

Solution: Given that the joint probability function

$$f(x, y) = \frac{1}{27}(2x + y);\ x = 0, 1, 2 \quad y = 0, 1, 2$$

The joint probability values are tabulated as follows:

$f(x, y)$ $x \rightarrow$ / $y \downarrow$	0	1	2
0	0	$\frac{2}{27}$	$\frac{4}{27}$
1	$\frac{1}{27}$	$\frac{3}{27}$	$\frac{5}{27}$
2	$\frac{2}{27}$	$\frac{4}{27}$	$\frac{6}{27}$

$$f(0,1) = \frac{1}{27}(2.0+1) = \frac{1}{27}$$

$$f(2,2) = \frac{1}{27}(2.2+1) = \frac{6}{27}$$

The marginal distribution of X is given by

$$P(X=0) = P(X=0,\ Y=0) + P(X=0,\ Y=1) + P(X=0,\ Y=2)$$

$$= 0 + \frac{1}{27} + \frac{2}{27}$$

$$= \frac{3}{27}$$

$$P(X=1) = P(X=1, Y=0) + P(X=1, Y=1) + P(X=1, Y=2)$$
$$= \frac{2}{27} + \frac{3}{27} + \frac{4}{27}$$
$$= \frac{9}{27}$$

$$P(X=2) = P(X=2, Y=0) + P(X=2, Y=1) + P(X=2, Y=2)$$
$$= \frac{4}{27} + \frac{5}{27} + \frac{6}{27}$$
$$= \frac{9}{27}$$

The distribution of X is given by

X	0	1	2	Total
$f_x(x)$	$\frac{3}{27}$	$\frac{9}{27}$	$\frac{15}{27}$	1

The conditional distribution for Y, given $X = x$ is

$$f_{\frac{y}{x}}\left(\frac{Y=y}{X=x}\right) = \frac{f(x,y)}{f_x(x)}$$

This can be tabulated as follows:

y \ x	0	1	2
0	0	$\frac{\frac{2}{27}}{\frac{9}{27}} = \frac{2}{9}$	$\frac{\frac{4}{27}}{\frac{15}{27}} = \frac{4}{15}$
1	$\frac{\frac{1}{27}}{\frac{3}{27}} = \frac{1}{3}$	$\frac{\frac{3}{27}}{\frac{9}{27}} = \frac{3}{9}$	$\frac{\frac{5}{27}}{\frac{15}{27}} = \frac{5}{15}$
2	$\frac{\frac{2}{27}}{\frac{3}{27}} = \frac{2}{3}$	$\frac{\frac{4}{27}}{\frac{9}{27}} = \frac{4}{9}$	$\frac{\frac{6}{27}}{\frac{15}{27}} = \frac{6}{15}$

EXAMPLE 2.11

Let X denote the reaction time in seconds, to a certain stimulant and Y denote the temperature in °F at which a certain reaction starts to take place. Suppose that the two random variables X and Y have the joint density,

$$f(x,y) = \begin{cases} 4xy, & 0 < x < 1,\ 0 < y < 1 \\ 0, & \text{otherwise} \end{cases}$$

(i) Find $P\left(0 \le X \le \frac{1}{2} \text{ and } \frac{1}{4} \le Y \le \frac{1}{2}\right)$.

(ii) Find $P(X < Y)$.

Solution:

(i) $$P\left(0 \le X \le \frac{1}{2}, \frac{1}{4} \le Y \le \frac{1}{2}\right) = \int_0^{\frac{1}{2}} \int_{\frac{1}{4}}^{\frac{1}{2}} f(x,y)\,dy\,dx$$

$$= \int_0^{\frac{1}{2}} \int_{\frac{1}{4}}^{\frac{1}{2}} 4xy\,dy\,dx = \int_0^{\frac{1}{2}} 4x \left|\frac{y^2}{2}\right|_{\frac{1}{4}}^{\frac{1}{2}} dx$$

$$= \int_0^{\frac{1}{2}} 2x \left|y^2\right|_{\frac{1}{4}}^{\frac{1}{2}} dx$$

$$= \int_0^{\frac{1}{2}} 2x \left[\frac{1}{4} - \frac{1}{16}\right] dx = \int_0^{\frac{1}{2}} 2x \left(\frac{4-1}{16}\right) dx$$

$$= \frac{2 \times 3}{16} \int_0^{\frac{1}{2}} x^2 dx = \frac{3}{8} \left|\frac{x^3}{3}\right|_0^{\frac{1}{2}} = \frac{3}{8}\left(\frac{1}{8 \times 3}\right)$$

$$P\left(0 \le X \le \frac{1}{2}, \frac{1}{4} \le Y \le \frac{1}{2}\right) = \frac{1}{64}$$

(ii) $$P(X < Y) = \int_{y=0}^{1} \left[\int_{x=0}^{Y} f(x,y)\,dx\right] dy$$

$$= \int_{y=0}^{1} \left[\int_{x=0}^{Y} 4xy\,dx\right] dy = \int_0^1 4y \left|\frac{x^2}{2}\right|_0^y dy$$

$$= \int_0^1 \frac{4y}{2} \left|x^2\right|_0^y = \int_0^1 2y(y^2) = 2\left|\frac{y^4}{4}\right|_0^1 = \frac{2}{4} = \frac{1}{2}$$

2.6 STOCHASTIC INDEPENDENCE

Let X and Y be two discrete or continuous random variables with joint pdf $f_{XY}(x, y)$ and with marginal pdf's $f_X(x)$, and $g_Y(y)$, respectively. If $g_Y\left(\frac{y}{x}\right)$ denotes the conditional pdf of Y given $X = x$, by compound probability theorem we have

$$f_{XY}(x, y) = f_X(x) \cdot g_Y\left(\frac{y}{x}\right) \tag{2.1}$$

If $g_Y\left(\frac{y}{x}\right)$ does not depend on x, then by definition of marginal density functions, we get for continuous random variables,

$$g_Y(y) = \int_{-\infty}^{\infty} f(x,y)\,dx$$

$$= \int_{-\infty}^{\infty} f_X(x) \cdot g\left(\frac{y}{x}\right) dx$$

$$= g\left(\frac{y}{x}\right) \int_{-\infty}^{\infty} f_X(x)\, dx$$

$$= g\left(\frac{y}{x}\right) \quad \left\{\text{since } g_Y\left(\frac{y}{x}\right) \text{ does not depend on } x.\right\}$$

$$\therefore g_Y(y) = g\left(\frac{y}{x}\right)$$

$$\therefore f_{XY}(x, y) = f_X(x) \cdot g_Y(y) \text{ from equation (2.1)}$$

Hence, two random variables X and Y are said to be stochastically independent if and only if

$$f(x, y) = f_X(x) \cdot g_Y(y)$$

In addition, $F(x, y) = F_X(x) \cdot G_Y(y)$ if F_X, G_Y denotes the distribution functions of X, Y and $F(x, y)$ denotes joint distribution function.

Worked Out Examples

EXAMPLE 2.12

Suppose the random variables X and Y have the point pdf $f(x,y) = \begin{cases} Kx(x-y), & 0 < x < 2, \\ & -x < y < x \\ 0, & \text{otherwise} \end{cases}$

(i) Determine the constant K.

(ii) Find marginal pdf of X and Y, respectively.

Solution:

(i) $f(x, y) = Kx(x - y)$

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)\, dx\, dy = 1$$

$$K \int_{x=0}^{2} \left[\int_{y=-x}^{x} (x^2 - xy)\, dy \right] dx = 1$$

$$K \int_{0}^{2} \left[x^2 |y|_{-x}^{x} - x \frac{y^2}{2}\Big|_{-x}^{x} \right] dx = 1$$

$$K \int_{0}^{2} \left[x^2(x + x) - x\left(\frac{x^2}{2} - \frac{x^2}{2}\right) \right] dx = 1$$

$$K \int_{0}^{2} 2x^3 dx = 1$$

$$2K \frac{x^4}{4}\Big|_{0}^{2} = 1$$

$$2K\left(\frac{16}{4}\right) = 1$$

$$8K = 1$$

$$K = \frac{1}{8}$$

Hence, the joint pdf is:

$$f(x,y) = \begin{cases} \frac{1}{8}x(x-y), & 0 < x < 2 \\ & -x < y < x \\ 0, & \text{otherwise} \end{cases}$$

(ii) The marginal pdf of X is given by

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)\,dy$$

$$= \frac{1}{8}\int_{-x}^{x} x(x-y)\,dy$$

$$= \frac{1}{8}\int_{-x}^{x} (x^2 - xy)\,dy = \frac{1}{8}\left[x^2 \left| y \right|_{-x}^{x} - x\left| \frac{y^2}{2} \right|_{-x}^{x} \right]$$

$$= \frac{1}{8}\left[2x^3 - x\left(\frac{x^2}{2} - \frac{x^2}{2} \right) \right]$$

$$f_X(x) = \begin{cases} \frac{x^3}{4}, & 0 < x < 2 \\ 0, & \text{otherwise} \end{cases}$$

The marginal pdf of Y is given by,

$$g_Y(y) = \int_{-\infty}^{\infty} f(x,y)\,dx$$

$$= \frac{1}{8}\int_{0}^{2} x(x-y)\,dx$$

$$= \frac{1}{8}\int_{0}^{2} (x^2 - xy)\,dx$$

$$= \frac{1}{8}\left[\left| \frac{x^3}{3} \right|_0^2 - y\left| \frac{x^2}{2} \right|_0^2 \right]$$

$$= \frac{1}{8}\left[\frac{8}{3} - y\left(\frac{4}{2} \right) \right] = \frac{1}{8}\left[\frac{8}{3} - 2y \right]$$

$$\therefore g_Y(y) = \begin{cases} \frac{1}{3} - \frac{y}{4}, & -x < y < x \\ 0, & \text{otherwise} \end{cases}$$

EXAMPLE 2.13

The joint probability density function of two dimensional random variable (X, Y) is given by,

$$f(x,y)=\begin{cases}\dfrac{x^3y^2}{16}, & 0\le x\le 2,\ 0\le y\le 2\\ 0, & \text{otherwise}\end{cases}$$

Find the marginal densities of X and Y. In addition, find the cumulative distributions for X and Y.

Solution: The marginal density of X is given by,

$$f_X(x)=\int_{-\infty}^{\infty} f(x,y)\,dy$$

$$=\int_0^2 \frac{x^3y^3}{16}dy=\frac{x^3}{16}\frac{y^4}{4}\Big|_0^2$$

$$f_X(x)=\begin{cases}\dfrac{x^3}{4}, & 0\le x\le 2\\ 0, & \text{otherwise}\end{cases}$$

The marginal density of Y is given by,

$$g_Y(y)=\int_{-\infty}^{\infty} f(x,y)\,dx$$

$$=\int_0^2 \frac{x^3y^3}{16}dx=\frac{y^3}{16}\frac{x^4}{4}\Big|_0^2$$

$$g_Y(y)=\begin{cases}\dfrac{y^3}{4}, & 0\le y\le 2\\ 0, & \text{otherwise}\end{cases}$$

The cumulative distribution function of X is given by,

$$F_X(x)=\int f_X(x)dx=\int\frac{x^3}{4}dx=\frac{x^4}{16}$$

$$\therefore F_X(x)=\begin{cases}0, & x<0\\ \dfrac{x^4}{16}, & 0\le x\le 2\\ 1, & x>2\end{cases}$$

Similarly, the cumulative distribution of Y is given by,

$$G_Y(y)=\int f_Y(y)dy=\int\frac{y^3}{4}dy$$

$$=\frac{y^4}{16}$$

$$\therefore G_Y(y) = \begin{cases} 0, & y < 0 \\ \dfrac{y^4}{16}, & 0 \le y \le 2 \\ 1, & y > 2 \end{cases}$$

EXAMPLE 2.14

A privately owned liquor store operates both a drive-in-facility and a walk-in-facility on a randomly selected day, let X and Y, respectively, be the proportions of the time that the drive-in and walk-in facilities are in use, the joint density function of these random variables is

$$f(x,y) = \begin{cases} \dfrac{2}{3}(x+2y), 0 \le x \le 1, 0 \le y \le 1 \\ 0, \qquad \text{otherwise} \end{cases}$$

(i) Find the marginal density of X.

(ii) Find the marginal density of Y.

(iii) Find the probability at the drive-in-facility is busy less than one half of the time.

Solution:

(i) The marginal density of X is given by,

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)dy$$

$$= \frac{2}{3}\int_0^1 (x+2y)dy$$

$$= \frac{2}{3}\left[x\left|y\right|_0^1 + 2\left|\frac{y^2}{2}\right|_0^1 \right]$$

$$f_X(x) = \begin{cases} \dfrac{2}{3}[x+1], & 0 \le x \le 1 \\ 0, \qquad \text{otherwise} \end{cases}$$

(ii) The marginal density of Y is given by,

$$g_Y(y) = \int_{-\infty}^{\infty} f(x,y)dx$$

$$= \frac{2}{3}\int_0^1 (x+2y)dx$$

$$= \frac{2}{3}\left[\left|\frac{x^2}{2}\right|_0^1 + 2y\left|x\right|_0^1 \right]$$

$$g_Y(y) = \begin{cases} \frac{2}{3}\left(\frac{1}{2} + 2y\right), & 0 \le y \le 1 \\ 0, & \text{otherwise} \end{cases}$$

(iii) Probability that the drive-in-facility is busy less than one half of the time is given by,

$$P\left(X < \frac{1}{2}\right) = \int_0^{\frac{1}{2}} f_X(x)\,dx$$

$$= \frac{2}{3}\int_0^{\frac{1}{2}} (x+1)\,dx$$

$$= \frac{2}{3}\left|\frac{x^2}{2} + x\right|_0^{\frac{1}{2}}$$

$$= \frac{2}{3}\left[\frac{1}{2}\left(\frac{1}{4} + \frac{1}{2}\right)\right]$$

$$= \frac{\cancel{2}}{\cancel{3}} \times \frac{1}{\cancel{2}} \times \frac{\cancel{3}}{4}$$

$$= \frac{1}{4}$$

EXAMPLE 2.15

Given the joint density function

$$f(x,y) = \begin{cases} \frac{6-x-y}{8}, & 0 < x < 2,\ 2 < y < 4 \\ 0, & \text{otherwise} \end{cases}$$

Find $P(2 < y < \frac{3}{X} = 2)$.

Solution: The marginal density of X is

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)\,dy$$

$$= \frac{1}{8}\int_2^4 (6-x-y)\,dy$$

$$= \frac{1}{8}\left[(6-x)\left|y\right|_2^4 - \left|\frac{y^2}{2}\right|_2^4\right]$$

$$= \frac{1}{8}[2(6-x) - (8-2)]$$

$$= \frac{1}{8}[12 - 2x - 6]$$

$$f_X(x) = \frac{1}{8}(6-2x),\ 0 < x < 2$$

$$f\left(\frac{y}{x}\right)=\frac{f(x,y)}{f_X(x)}=\frac{\frac{1}{8}(6-x-y)}{\frac{1}{8}(6-2x)}$$

$$=\frac{6-x-y}{6-2x}$$

$$P\left(a<Y<\frac{b}{X}=x\right)=\int_a^b f\left(\frac{y}{x}\right)dy$$

$$P\left(2<Y<\frac{3}{X}=2\right)=\int_2^3 f\left(\frac{y}{x}\right)dy$$

$$=\int_2^3 \frac{6-x-y}{6-2x}\bigg|_{x=2} dy$$

Denominator is

$$\int_2^3 (6-2x)\,dx \text{ for } x=2$$

Denominator = 2.

$$=\int_2^3 \frac{6-2-y}{6-4}dy$$

$$=\frac{1}{2}\int_2^3 (4-y)\,dy$$

$$=\frac{1}{2}\left[4|y|_2^3-\left|\frac{y^2}{2}\right|_2^3\right]=\frac{1}{2}\left[4-\frac{1}{2}(5)\right]$$

$$=\frac{1}{2}\left[\frac{3}{2}\right]$$

$$\therefore P\left(2<Y<\frac{3}{X}=2\right)=\frac{3}{4}$$

EXAMPLE 2.16

Given the joint density of X and Y is

$$f(x,y)=\begin{cases}6x, & 0<x<2, 0<y<1-x\\ 0, & \text{elsewhere}\end{cases}$$

(i) Show that X and Y are not independent.

(ii) Find $\mathrm{P}\left(\frac{X>0.3}{Y=0.5}\right)$.

Solution:

(i) Two random variables X and Y are said to be stochastically independent if

$$f(x, y)=f_X(x)\times f_Y(y)$$

The marginal density of X is

$$f_x(x) = \int_{-\infty}^{\infty} f(x,y)\,dy = 6\int_0^{1-x} x\,dy$$

$$f_x(x) = 6x\,|y|_0^{1-x}$$

$$= 6x(1-x), \quad 0 < x < 2$$

The marginal density of Y is

$$f_y(y) = \int_{-\infty}^{\infty} f(x,y)\,dx$$

$$= \int_0^1 6x\,dx = 6\frac{x^2}{2}\Big|_0^2$$

$$f_y(y) = \begin{cases} 12, & 0 < y < 1-x \\ 0, & \text{otherwise} \end{cases}$$

$$f_x(x) \times f_y(y) = 6x(1-x) \times 12$$

$$= 72x(1-x)$$

$$\neq f(x,y)$$

$\therefore$ X and Y are not stochastically independent.

(ii) $$P\left(\frac{X > 0.3}{Y = 0.5}\right) = \int_{0.3}^{2} f\left(\frac{x}{y}\right)dx$$

$$f\left(\frac{x}{y}\right) = \frac{f(x,y)}{f_Y(y)} = \frac{6x}{12} = \frac{x}{2}$$

$$P\left(\frac{X > 0.3}{Y = 0.5}\right) = \int_{0.3}^{2} \frac{x}{2}\Big|_{y=0.5}\,dx$$

$$= \frac{1}{2}\left|\frac{x^2}{2}\right|_{0.3}^{2}$$

$$= \frac{1}{4}[4 - 0.09]$$

$$= \frac{3.91}{4}$$

$$= 0.9775.$$

EXAMPLE 2.17

From a bag of fruits which contain 4 oranges, 2 apples, and 2 bananas, a random sample of 4 fruits are selected. If X is the number of oranges and Y is the number of apples, in the sample then:

(i) Find the joint distribution of X and Y.

(ii) $P[(X, Y)A]$ where A is the region $\left\{\frac{(x,y)}{x} + y \leq 2\right\}$.

(iii) Find the marginal densities of X, Y.

Solution: The possible pairs of values (x, y) are (0, 0), (0, 1), (1, 0), (1, 1), (0, 2), (2, 0).

The number of ways of selecting 1 orange from 4 oranges and 1 apple from 2 apples is $(4C_1)(2C_1) = 8$ ways.

The total number of equally likely ways of selecting any two fruits from 8 fruits is $8C_2 = 28$ ways.

$\therefore f(1, 1)$ represents the probability that one orange and one apple are selected.

$$\therefore f(1,1) = \frac{8}{28} = \frac{4}{14} = \frac{2}{7}$$

Hence, the joint probability distribution of (X, Y) can be represented by the formula

$$f(x,y) = \frac{(4C_x)(2C_y)(2C_{2-x-y})}{(8C_2)}$$

The joint probability distribution of X and Y is given in the following table:

$f(x, y)$		x			Total $g(y)$
		0	1	2	
y	0	$\frac{1}{28}$	$\frac{8}{28}$	$\frac{6}{28}$	$\frac{15}{28}$
	1	$\frac{4}{28}$	$\frac{8}{28}$	–	$\frac{12}{28}$
	2	$\frac{1}{28}$	–	–	$\frac{1}{28}$
Total $f(x)$		$\frac{6}{28}$	$\frac{16}{28}$	$\frac{6}{28}$	1

Since the marginal densities of X are given by,

x	0	1	2
$f(x)$	$\frac{6}{28}$	$\frac{16}{28}$	$\frac{6}{28}$

Similarly, marginal density of Y is given by,

y	0	1	2
$g(y)$	$\frac{15}{28}$	$\frac{12}{28}$	$\frac{1}{28}$

From the above two tables it is clear that

$$\sum_x f(x) = \frac{6}{28} + \frac{16}{28} + \frac{6}{28} = 1$$

$$\sum_y g(y) = \frac{15}{28} + \frac{12}{28} + \frac{1}{28} = 1$$

EXAMPLE 2.18

Given $f(x, y) = e^{-(x+y)}$, $0 \le x < \infty$, $0 \le y < \infty$

(i) Find $P(X > 1)$

(ii) Find $P\left(X < \frac{Y}{X} < 2Y\right)$

(iii) Find $P(X + Y < 1)$

Solution:

(i) The marginal density of X is given by

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)\,dy = \int_0^{\infty} e^{-x-y}\,dy$$

$$= e^{-x}\left|\frac{e^{-y}}{-1}\right|_0^{\infty} = e^{-x}, \quad 0 \le x < \infty$$

Hence,

$$P(X > 1) = \int_1^{\infty} f_X(x)\,dx$$

$$= \int_1^{\infty} e^{-x}\,dx = \left.\frac{e^{-x}}{-1}\right|_1^{\infty}$$

$$= e^{-1}$$

(ii) $$P\left(X < \frac{Y}{X} < 2Y\right) = \frac{P(X < Y \cap X < 2Y)}{P(X < 2Y)}$$

$$= \frac{P(X < Y)}{P(X < 2Y)}$$

$$P(X < Y) = \int_0^{\infty}\left[\int_0^{y} f(x,y)\,dx\right]dy$$

$$= \int_0^{\infty}\left[\int_0^{y} e^{-x-y}\,dx\right]dy$$

$$= \int_0^{\infty}\left[e^{-y}\int_0^{y} e^{-x}\,dx\right]dy$$

$$= \int_0^{\infty} e^{-y}\left.\frac{e^{-x}}{-1}\right|_0^{y} dy$$

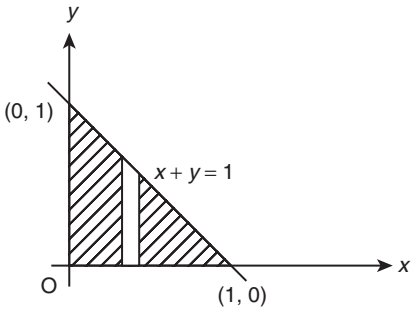
$$= \int_0^{\infty} e^{-y}(1 - e^{-y})\,dy = \int_0^{\infty}(e^{-y} - e^{-2y})\,dy$$

$$= \left.\frac{e^{-y}}{-1}\right|_0^{\infty} - \left.\frac{e^{-2y}}{-2}\right|_0^{\infty}$$

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

$$P(X + Y < 1) = \iint f(x, y)\, dx dy$$

$$\therefore P(X + Y < 1) = \int_{x=0}^{1} \left[\int_{y=0}^{1-x} f(x, y)\, dy \right] dx$$



$$= \int_0^1 e^{-x} \left[\int_{y=0}^{1-x} e^{-y} dy \right] dx$$

$$= \int_0^1 e^{-x} \left| \frac{e^{-y}}{-1} \right|_0^{1-x} dx = \int_0^1 e^{-x} (-e^{-(1-x)} + 1) dx$$

$$= \int_0^1 (-e^{-x-1+x} + e^{-x}) dx = \int_0^1 (e^{-1} + e^{-x}) dx$$

$$= -e^{-1} |x|_0^1 + \left| \frac{e^{-x}}{-1} \right|_0^1 = -e^{-1} + 1 - e^{-1}$$

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

EXAMPLE 2.19

Let X and Y be random variables with joint distribution given as follows:

x \ y	−3	2	4
1	0.1	0.2	0.2
3	0.3	0.1	0.1

(i) Find the marginal distributions of X and Y.

(ii) Find whether X and Y are independent random variables.

Solution: The marginal distributions of X and Y are obtained as follows:

x \ y	−3	2	4	$f(x)$
1	0.1	0.2	0.2	0.5
3	0.3	0.1	0.1	0.5
$g(y)$	0.4	0.3	0.3	1

Marginal density of X

x	1	3
$f(x)$	0.5	0.5

It is clear that $\sum_x f(x) = 0.5 + 0.5$
$= 1$

Marginal density of Y

y	−3	2	4
$g(y)$	0.4	0.3	0.3

$\sum_y g(y) = 0.4 + 0.3 + 0.3$
$= 1$

To check stochastic independence
$f(1, -3) = 0.1, f(1) = 0.5$, and $g(-3) = 0.4$
$f(1, -3) = f(1)\, g(-3)$
$0.1 \neq (0.5)\ (0.4)$
Hence, X and Y are not stochastically independent.

EXAMPLE 2.20

Let X denote the number of times a certain numerical control machine will malfunction: 1, 2, or 3 times on any given day. Let Y denote the number of times a technician is called on an emergency call. This joint probability distribution is given as:

$f(x, y)$		x		
		1	2	3
	1	0.05	0.05	0.1
y	2	0.05	0.1	0.35
	3	0	0.2	0.1

(i) Evaluate the marginal distribution of X.

(ii) Evaluate marginal distribution of Y.

(iii) Find $P\left(\frac{Y=1}{X=2}\right)$, $P\left(\frac{Y=2}{X=3}\right)$, $P\left(\frac{Y=3}{X=1}\right)$, $P\left(\frac{X=1}{Y=2}\right)$, $P\left(\frac{X=2}{Y=3}\right)$ and $P\left(\frac{X=3}{Y=1}\right)$

Solution: Consider the joint probability distribution:

$f(x, y)$		x			Total $g(y)$
		0	1	2	
	1	0.05	0.05	0.1	0.2
y	2	0.05	0.1	0.35	0.5
	3	0	0.2	0.1	0.3
	Total $f(x)$	0.1	0.35	0.55	1

(i) The marginal density of X is given by

x	1	2	3
$f(x)$	0.1	0.35	0.55

It is clear that $\sum_x f(x) = 0.1 + 0.35 + 0.35 = 1$

(ii) The marginal density of Y is given by

y	1	2	3
$g(y)$	0.2	0.5	0.3

$$\sum_y g(y) = 0.2 + 0.5 + 0.3 = 1$$

(iii) $$P\left(\frac{Y=1}{X=2}\right) = \frac{P(Y=1, X=2)}{P(X=2)} = \frac{0.05}{0.35} = 0.143$$

$$P\left(\frac{Y=2}{X=3}\right) = \frac{P(X=3, Y=2)}{P(X=3)} = \frac{0.35}{0.55} = 0.636$$

$$P\left(\frac{Y=3}{X=1}\right) = \frac{P(X=1, Y=3)}{P(X=1)} = \frac{0}{0.1} = 0$$

$$P\left(\frac{X=1}{Y=2}\right) = \frac{P(X=1, Y=2)}{P(Y=2)} = \frac{0.05}{0.5} = 0.1$$

$$P\left(\frac{X=2}{Y=3}\right) = \frac{P(X=2, Y=3)}{P(Y=3)} = \frac{0.2}{0.3} = 0.66$$

$$P\left(\frac{X=3}{Y=1}\right) = \frac{P(X=3, Y=1)}{P(Y=1)} = \frac{0.1}{0.2} = 0.5$$

Work Book Exercises

6. Explain the following concepts:
   (i) Marginal density functions.
   (ii) Conditional distributions.
   (iii) Independence of random variables.

7. The joint probability distribution of a pair of random variables is given by the following table:

y \ x	1	2	3
1	0.25	0.05	0.25
2	0.35	0.05	0.05

(i) Find the marginal density functions of X and Y.

(ii) Find the conditional distribution of X given $Y = 1$.

(iii) Find $P[(X + Y) < 4]$.

8. The joint probability of two discrete random variables X and Y is as follows:

y \ x	1	2	3
1	$\frac{3}{32}$	$\frac{5}{32}$	$\frac{1}{32}$
2	$\frac{11}{32}$	$\frac{7}{32}$	$\frac{3}{32}$
3	$\frac{1}{32}$	0	$\frac{1}{32}$

(i) Find the marginal distributions of X and Y.

(ii) Evaluate the conditional distribution of X given $Y = 2$ and conditional distribution of Y given $X = 2$.

9. Two discrete random variables X and Y have

$P(X = 0, Y = 0) = \frac{2}{9}, P(X = 0, Y = 1) = \frac{1}{9}$

$P(X = 1, Y = 0) = \frac{1}{9}, P(X = 1, Y = 1) = \frac{5}{9}$. Examine whether X and Y are independent.

10. An urn contains 3 blue balls, 2 red balls, and 3 orange balls. Two balls are selected at random. If X is the number of blue balls and Y is number of red balls selected then

(i) Find the joint probability function $f(x, y)$.

(ii) Find $P\left[\frac{(X, Y)}{A}\right]$ where A is the region $\left\{\frac{(x, y)}{x} + y \leq 1\right\}$

(iii) Find the marginal distributions of X and Y.

2.7 TRANSFORMATION OF ONE-DIMENSIONAL RANDOM VARIABLE

Sometimes it is required to derive the probability distribution of a function of one or more random variables.

Let X be a random variable defined on the sample space S and let $Y = g(X)$ be a random variable defined on S. Suppose $f_X(x)$ is the probability density function of X, then the density function of Y is given by,

$$h_Y(y) = f_X(x)\left|\frac{dx}{dy}\right|$$

where x is expressed in terms of y.

If $F_X(x)$ denotes the cumulative distribution function of X, then the cumulative distribution function of Y is

$H_Y(y) = F_X(x)$ when $Y = y(x)$ is a strictly increasing function and incase strictly decreasing function,

$$H_Y(y) = 1 - F_X(x)$$

Worked Out Examples

EXAMPLE 2.21

Find the pdf of $Y = 2X^2 - 3$ given the pdf of X as $f(x) = \begin{cases} \frac{1}{6}, & -3 \le x \le 3 \\ 0, & \text{elsewhere} \end{cases}$

Solution: From the definition of transformation of random variable,

$$h_Y(y) = f_X(x)\left|\frac{dx}{dy}\right|$$

$$= \frac{1}{6}\cdot\frac{dx}{dy}$$

$$\because Y = 2X^2 - 3$$

$$x = \left(\frac{y+3}{2}\right)^{\frac{1}{2}}$$

$$\frac{dx}{dy} = \frac{1}{2}\left(\frac{y+3}{2}\right)^{\frac{-1}{2}}\cdot\frac{1}{2}$$

$$= \frac{1}{4}\left(\frac{y+3}{2}\right)^{\frac{-1}{2}}$$

$$\therefore h_Y(y) = \frac{1}{6}\cdot\frac{1}{4}\left(\frac{y+3}{2}\right)^{\frac{-1}{2}}, \quad -3 \le x \le 3$$

$$|x| \le 3$$

$$\left|\frac{y+3}{2}\right|^{\frac{1}{2}} \le 3$$

$$|y| \le 15$$

$$\therefore h_Y(y) = \begin{cases} \frac{1}{24}\frac{\sqrt{2}}{\sqrt{y+3}}, & |y| \le 15 \\ 0, & \text{elsewhere} \end{cases}$$

EXAMPLE 2.22

Find the probability distribution of Y where $Y = 8X^3$ given the random variable X with probability distribution:

$$f(x) = \begin{cases} 2x, & 0 < x < 1 \\ 0, & \text{elsewhere} \end{cases}$$

Solution: From the definition of transformation of a random variable,

$$h_Y(y) = f_X(x)\left|\frac{dx}{dy}\right|$$

Given $y = 8x^3$,

$$\frac{dx}{dy} = \frac{d}{dy}\left(\frac{y}{8}\right)^{\frac{1}{3}} = \frac{1}{2}\frac{1}{3}y^{\frac{1}{3}-1}$$

$$= \frac{1}{6y^{\frac{2}{3}}}$$

$$\therefore h_Y(y) = 2x \cdot \frac{1}{6y^{\frac{2}{3}}}$$

$$= \frac{2}{2}y^{\frac{1}{3}} \cdot \frac{1}{6y^{\frac{2}{3}}}$$

$$= \frac{1}{6y^{\frac{1}{3}}}, 0 < x < 1$$

$$0 < \frac{y^{\frac{1}{3}}}{2} < 1$$

$$0 < y^{\frac{1}{3}} < 2$$

$$0 < y < 8$$

The pdf of Y is

$$\therefore h_Y(y) = \begin{cases} \frac{1}{6y^{\frac{1}{3}}}, & 0 < y < \infty \\ 0, & \text{otherwise} \end{cases}$$

EXAMPLE 2.23

Find the density function and the distribution function of $Y = X^3$ given the density function of X is $f(x) = e^{-x}$, $0 \le x \le \infty$.

Solution: From the definition of transformation of random variables, we have

$$h_Y(y) = f_X(x)\left|\frac{dx}{dy}\right|$$

$$\because y = x^3 \Rightarrow x = y^{\frac{1}{3}}$$

$$\frac{dx}{dy} = \frac{1}{3}y^{\frac{-2}{3}}$$

$$h_Y(y) = e^{-x} \cdot \frac{1}{3}y^{\frac{-2}{3}}$$

$$= e^{-y^{\frac{1}{3}}} \cdot \frac{1}{3}y^{\frac{2}{3}},\ 0 < x < \infty$$

$$0 \le y^{\frac{1}{3}} < \infty$$

$\therefore$ The pdf of Y is

$$h_Y(y) = \begin{cases} \frac{1}{3}e^{-y^{\frac{1}{3}}}y^{\frac{2}{3}}, & 0 < y \le \infty \\ 0, & \text{elsewhere} \end{cases}$$

The cdf of $X = P(X = x) = \int_0^x f(x)\,dx = \int_0^x e^{-x}dx$

$$= \frac{e^{-x}}{1}\Big|_0^x = 1 - e^{-x}$$

The cdf of Y is given by

$$H_Y(y) = 1 - F_X(x)$$

$$= 1 - [1 - e^{-x}] \text{ since } y = g(x) \text{ is strictly decreasing}$$

$$= e^{-x}$$

$$H_Y(y) = e^{-y^{\frac{1}{3}}}$$

EXAMPLE 2.24

Find the distribution and density functions of the random variable

(i) $Y = a + bX$,

(ii) $Y = \cos X$ given the density function of X as

$$f(x) = \begin{cases} \frac{1}{2}, & -1 < x < 1 \\ 0, & \text{elsewhere} \end{cases}$$

Solution:

(i) $Y = a + bX$

$$x = \frac{y - a}{b}$$

$$\frac{dx}{dy} = \frac{1}{b}$$

The pdf of Y is

$$h_Y(y) = f_X(x)\left|\frac{dx}{dy}\right|$$

$$= \frac{1}{2}\cdot\frac{1}{b} \quad \begin{matrix} -1 < x < 1 \\ -1 < \dfrac{y-a}{b} < 1 \end{matrix}$$

$$\therefore h_Y(y) = \begin{cases} \dfrac{1}{2b}, a-b < y < a+b & -b < y-a < b \\ 0, \quad \text{elsewhere} & a-b < y < a+b \end{cases}$$

The distribution function of Y is given by

$$H_Y(y) = F_X(x)$$

where the distribution of X is given by

$$F_X(x) = P(X \le x) = \int_{-1}^{x} f(x)\,dx$$

$$= \frac{1}{2}\int_{-1}^{x} dx = \frac{x}{2}\Big|_{-1}^{x}$$

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

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

$$\therefore H_Y(y) = \frac{1}{2}(x+1)$$

$$= \frac{1}{2}\left(\frac{y-a}{b}+1\right)$$

$$= \frac{1}{2b}(y-a+b)$$

(ii) The density function of Y is

$$h_Y(y) = f_X(x)\left|\frac{dx}{dy}\right|$$

$\because y = \cos x \Rightarrow x = \cos^{-1} y$

$$\frac{dx}{dy} = \frac{-1}{\sqrt{1-y^2}}$$

$$\therefore h_Y(y) = \frac{-1}{2}\frac{1}{\sqrt{1-y^2}} \quad \begin{matrix} -1 \le x \le 1 \\ -1 \le \cos^{-1} y \le 1 \end{matrix}$$

So far we have seen the joint probability distributions when the transformation is to a single random variable. However, let us look at transformation of both the random variables X and Y to other variables namely U and V, respectively.

2.8 TRANSFORMATION OF TWO-DIMENSIONAL RANDOM VARIABLE

Let the random variables X and Y be transformed to two random variables U and V. Let the transformations be $u = u(x, y)$ and $v = v(x, y)$ where, u and v are continuously differentiable functions. The joint pdf of the transformed variables U and V is given by

$$g_{uv}(u,v) = f_{XY}(x,y) \cdot |J|$$

where $|J|$ is the modulus of Jacobian of x and y, w.r.t u and v and $f(x, y)$ is expressed in terms of u and v.

The Jacobian of the transformations is obtained by

$$J = \frac{\partial(x,y)}{\partial(u,v)} = \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{vmatrix} > 0 \text{ or } < 0.$$

The inverse functions are uniquely given by

$$x = x(u, v), y = y(u, v).$$

Worked Out Examples

EXAMPLE 2.25

Let X and Y be two continuous random variables with joint probability distribution,

$$f(x,y) = \begin{cases} 4xy, & 0 < x < 1, 0 < y < 1 \\ 0, & \text{otherwise} \end{cases}$$

Find the joint probability distribution of $U = X^2$, and $V = XY$.

Solution: The inverse solutions of $U = X^2$ and $V = XY$ are $x = \sqrt{U}$, $y = \dfrac{v}{\sqrt{U}}$

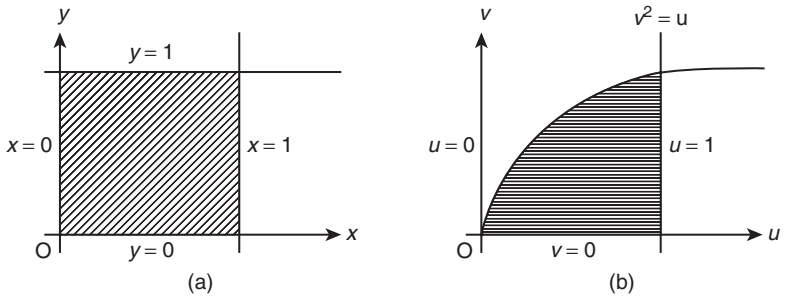
Hence, the Jacobian of x, y w.r.t u, v is

$$J = \frac{\partial(x,y)}{\partial(u,v)} = \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{vmatrix}$$

$$= \begin{vmatrix} \frac{1}{2\sqrt{u}} & 0 \\ v\left(\frac{-1}{2}\right)u^{\frac{-3}{2}} & u^{\frac{-1}{2}} \end{vmatrix}$$

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

The regions which can be obtained from *xy* plane into *uv* plane are as follows:
$x = 0, y = 0$ and $x = 1, y = 1$ are transformed to $u = 0$, $u = 1$, $v = 0$, $v = \sqrt{u}$, $v^2 = u$.



The joint distribution of (u, v) is given as,

$$g_{uv}(u,v) = f_{XY}(x,y)|J|$$

$$= 4xy \cdot \frac{1}{2u}$$

$$= 4(\sqrt{u})\left(\frac{v}{\sqrt{u}}\right)\cdot\frac{1}{2u}$$

$$= \begin{cases} \dfrac{2v}{u}, v^2 < u < 1, 0 < v < 1 \\ 0, \quad \text{elsewhere} \end{cases}$$

Work Book Exercises

11. Let (X, Y) be a two dimensional non-negative continuous random variables having the joint density

$$f(x,y) = \begin{cases} 4xy\, e^{-(x^2+y^2)}, & x \geq 0, y \geq 0 \\ 0, & \text{elsewhere} \end{cases}$$

Prove that the density function of $U = \sqrt{x^2 + y^2}$ is

$$h(u) = \begin{cases} 2u^3 e^{-u^2}, 0 \leq u < \infty \\ 0, \quad \text{elsewhere} \end{cases}$$

DEFINITIONS AT A GLANCE

Random Variable: A random variable on a sample space is a rule that assigns a numerical value to each outcomes of *S*.

Discrete Random Variable: A random variable which assumes only integral values in an interval of domain is discrete random variable.

Continuous Random Variable: Quantities which are capable of taking all values in a range are continuous random variables.

Probability Mass Function: $P(X = x)$ is called pmf, if $P(X = x) \geq 0$ and $\Sigma P(X = x) = 1$.

Probability Density Function: $f(x)$ is called pdf, if it satisfies $f(x) \geq 0$ and $\int_a^b f(x)\,dx = 1$.

Cumulative Distribution Function: The cdf, of $f(x)$ denoted by $F(x)$ is given by $F(x) = \int_{-\infty}^{x} f(x)\,dx$

Joint Probability Distribution: If X, Y are discrete random variables then $P(X = x_i, Y = y_j)$ is the joint probability distribution.

Marginal Density Function: Marginal probability function of X is $P_{ij} = \sum_{i=1}^{m} P(x_i, y_j)$.

Stochastic Independence: Two random variables X and Y are said to be stochastically independent if and only if $f(x, y) = f_X(x) \cdot g_Y(y)$.

FORMULAE AT A GLANCE

- $P(X = x)$ is called pdf/pmf if
  - $P(X = x) \geq 0$
  - $\Sigma P(X = x) = 1$ for a discrete variable
  - $P(X = x) \geq f(x)$ is pdf if $P(X = x) \geq 0$ and $\int_{-\infty}^{\infty} f(x)dx = 1$
- $P(X = x_i, Y = y_j)$ is called joint probability mass function if
  - $\sum_{i=1}^{m}\sum_{j=1}^{n} P(X = x_i, Y = y_j) = P(x_i, y_j) = 1$
  - $P(X = x_i, Y = y_j) = P(X = x_i \cap Y = y_j)$
- The marginal probability function of X is

  $$p_{\cdot j} = \sum_{i=1}^{m} p(x_i, y_j)$$

  In addition, $\sum_{j=1}^{n} p_{\cdot j} = 1$
- The marginal probability function of Y is

  $$p_{i\cdot} = \sum_{j=1}^{n} p(x_i, y_j)$$

  In addition, $\sum_{i=1}^{m} p_{i\cdot} = 1$
- For continuous random variables X and Y, the joint density function $f(x, y)$ must satisfy the following:
  - $f(x, y) \geq 0 \ \forall \ (x, y)$
  - $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x, y)\,dx\,dy = 1$
- Two random variables X, Y are stochastically independent if $f_{XY}(x, y) = f_X(x) \times g_Y(y)$

- By transforming a random variable $Y = g(X)$, if $f_X(x)$ is the probability density function of X, then the probability density function of Y is

$$h_Y(y) = f_X(x)\left|\frac{dx}{dy}\right|$$

- By transforming two dimensional random variables X and Y to u, v where $u = u(x, y)$ and $v = v(x, y)$, then the joint pdf of transformed variables u, v is given by

$$g_{uv}(u,v) = f_{XY}(x,y)\cdot|J|$$

where J is the Jacobian of (x, y) w.r.t (u, v) given by

$$J = \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{vmatrix}$$

OBJECTIVE TYPE QUESTIONS

1. The value of k for a valid probability density function for $f(x) = kx^2$, $0 < x < 1$ is _____.
   (a) 2 (b) 1
   (c) 3 (d) 0

2. If $f(x) = kx(1 - x)$ in $0 < x < 1$ is a probability density function, then the value of k is _____.
   (a) 6 (b) 5
   (c) 4 (d) $\frac{1}{6}$

3. The probability density function $f(x)$ for a continuous random variable X is defined by $f(x) = \begin{cases} \frac{k}{x^2}, & 5 \le x \le 10 \\ 0, & \text{elsewhere} \end{cases}$. Then the value of k is _____.
   (a) $\frac{1}{10}$ (b) 10
   (c) 5 (d) $\frac{1}{5}$

4. The probability density function $f(x) = \frac{x}{8}$, $0 \le x \le 4$, the mean is _____.
   (a) $\frac{8}{3}$ (b) $\frac{5}{3}$
   (c) $\frac{7}{3}$ (d) $\frac{1}{3}$

5. If the probability density function $f(x) = \frac{x^2}{9}$, $0 \le x \le 3$, the mean of the distribution is _____.
   (a) $\frac{1}{12}$ (b) $\frac{25}{12}$
   (c) $\frac{23}{12}$ (d) $\frac{27}{12}$

6. The relation between probability density function and cumulative distribution function of a random variable is _____.

   (a) $f(x) = \frac{d}{dx}F(x)$ (b) $F(x) = \frac{d}{dx}f(x)$

   (c) $f(x) = \int F(x)\,dx$ (d) No relation

7. If the distribution function for X is $F(x) = \begin{cases} 0, & x < 0 \\ 1 - \frac{1}{4}e^{-x}, & \text{for } x \geq 0 \end{cases}$, then $P(X = 0)$ is _____.

   (a) $\frac{3}{4}$ (b) $\frac{1}{2}$

   (c) $\frac{1}{4}$ (d) 0

8. If $f(x) = (1 - k)\, e^{-x}$, $x = 0, 1, 2, 3,\ldots$ is a probability distribution of a random variable, then $k =$ _____.

   (a) 1 (b) $1 + e$

   (c) $1 - e$ (d) e

9. If X is a discrete random variable whose probability distribution is $f(x) = -x$, $x = 1, 2, 3, 4, 5$, then the value of $k =$ _____.

   (a) $\frac{5}{8}$ (b) $\frac{3}{8}$

   (c) $\frac{7}{8}$ (d) $\frac{1}{8}$

10. The total probability under a density curve is _____.

    (a) 1 (b) 0

    (c) −1 (d) ∞

11. If $f(x)$ is a valid probability density function given by $f(x) = \begin{cases} kx, & 0 \leq x \leq 5 \\ k(10 - x), & 5 \leq x \leq 10 \\ 0, & \text{otherwise} \end{cases}$, then the value of k is _____.

    (a) $\frac{5}{25}$ (b) $\frac{3}{25}$

    (c) $\frac{1}{25}$ (d) $\frac{2}{5}$

ANSWERS

1. (c) 2. (a) 3. (b) 4. (a) 5. (d) 6. (a) 7. (c) 8. (c)
9. (a) 10. (a) 11. (c)

3 Mathematical Expectation

Prerequisites

Before you start reading this unit, you should:

- Be able to differentiate between discrete and continuous random variables
- Be familiar with pmf and pdf
- Calculate marginal density of random variables

Learning Objectives

After going through this unit, you would be able to

- Understand the expectation, variance, conditional mean, Chebychev's theorem, moments, moment generating function, and characteristic function
- Calculate expectation, variance, conditional mean, given joint density functions of X and Y
- Apply Chebychev's theorem to a random variable with some mean and variance

INTRODUCTION

Consider for instance, a salesman's average monthly income. This average may not be equal to any of his monthly pay. As an another example, if two coins are tossed 10 times and if X is the number of heads that occurs per toss, then the value of X can be 0 or 1 or 2. Suppose these have occurred say 3, 5, and 2 times, respectively. Then the average number of heads per toss of the two coins can be obtained as:

$$\frac{0(3)+1(5)+2(2)}{10}=0.9$$

This is the average value and need not be a possible outcome of the experiment. This average value is referred to as mean of the random variable X denoted by μ. Hence, we should know the values of the random variable and their relative frequencies. This is also referred to as mathematical expectation or the expected value of the random variable X.

Worked Out Examples

EXAMPLE 3.1

When a car is washed, an attendant is paid according to the number of cars that pass through. Suppose the probabilities are $\frac{1}{12},\frac{1}{12},\frac{1}{4},\frac{1}{4},\frac{1}{6}$, and $\frac{1}{6}$, respectively that the attendant receives ₹70, ₹90, ₹110, ₹130, ₹150, and ₹170 on any sunny day. Find the attendant's expected earnings for this particular day.

Solution: Let X be a random variable, which assumes the attendant receiving a particular amount with the given probabilities. It can be given as a distribution.

The probability distribution of X is given by

$X = x$	70	90	110	130	150	170
$P(X = x)$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{6}$	$\frac{1}{6}$

The expected earnings of the attendant are obtained using,

$$E(X) = \sum_{x=1}^{6} xP(X = x)$$

$$= 70\left(\frac{1}{12}\right) + 90\left(\frac{1}{12}\right) + 110\left(\frac{1}{4}\right) + 130\left(\frac{1}{4}\right) + \frac{150}{6} + \frac{170}{6}$$

$$= 5.833 + 7.5 + 27.5 + 32.5 + 25 + 28.33$$

$$= 126.66$$

Hence, the attendant's expected earnings for this particular day is ₹126.66.

EXAMPLE 3.2

The probability distribution of the discrete random variable X is $f(x) = {}_3C_x \left(\frac{1}{4}\right)^x \left(\frac{3}{4}\right)^{3-x}$, $x = 0, 1, 2, 3$. Find the mean of X.

Solution: The probability distribution of X is given by

$X = x$	0	1	2	3
$P(X = x)$	$\left(\frac{3}{4}\right)^3$	$\left(\frac{3}{4}\right)^3$	$\left(\frac{1}{4}\right)\left(\frac{3}{4}\right)^2$	$\left(\frac{1}{4}\right)^3$

The mean of X is given by

$$E(X) = \sum_{x=0}^{3} xP(X = x)$$

$$= 0\left(\frac{3}{4}\right)^3 + 1\left(\frac{3}{4}\right)^3 + \frac{2}{4}\left(\frac{3}{4}\right)^2 + 3\left(\frac{1}{4}\right)^3$$

$$= 0.421875 + 0.28125 + 0.046875$$

$$= 0.75$$

$\therefore$ Mean of $X = 0.75$.

3.1 MATHEMATICAL EXPECTATION

Let X be a discrete random variable with pmf $P(X = x)$. Then its mathematical expectations denoted by $E(X)$ is given by

$$\mu = E(X) = \sum_{x} xP(X = x)$$

If X is a continuous random variable with pdf $f(x)$, then its expectation is given by

$E(X) = \int_{-\infty}^{\infty} x f(x)\, dx$. This is also called mean of X.

Properties of Expectation

1. If X and Y are random variables then $E(X + Y) = E(X) + E(Y)$ provided all the expectations exist.
2. The mathematical expectation of the sum of n random variables is equal to the sum of their expectations, provided all the expectations exist.
3. If X and Y are independent random variables, then $E(XY) = E(X) \cdot E(Y)$.
4. The mathematical expectation of the product of n independent random variables is equal to the product of their expectations.
5. If X is a random variable and 'a' is a constant then

$$E[af(X)] = aE[f(X)] \text{ and}$$

$$E[f(X) + a] = E[f(X)] + a$$

6. If X is a random variable and a, b are constants then $E(aX + b) = aE(X) + b$ provided all the expectations exist.

Convex Function

A real valued function f: $X \to R$ is called convex if, for any two points x_1 and x_2 in X and any $t \in [0,1]$,

$$f[tx_1 + (1 - t)x_2] \leq tf(x_1) + (1 - t)f(x_2)$$

Jensen's Inequality

If X is a random variable taking values in the domain of f, then $E[f(X)] \geq f[E(X)]$.

3.2 VARIANCE

In the earlier topic, we have studied about mean of a random variable X, which is very important in statistics because it describes where the probability distribution is centred. However, the mean itself is not adequate enough to depict the shape of the distribution, it deals with. Hence, we need to characterize the variability in the distribution.

Consider the following two classes of observations which give the same averages:

I:	61	43	34	46	55	45	50	40	39	47	57
II:	39	30	22	33	75	45	72	19	52	60	70

The average of both classes $= \frac{517}{11} = 47.$

In both the classes, we observe that the averages of both the classes is the same, but when the observations are observed, the first distribution has all its values centred around the average value, whereas the second distribution has all its values scattered away from the average value. Hence, the mean does not give the distribution completely. Therefore, variance of a random variable is another important measure of a distribution.

Variance of X

Let X be a discrete random variable. The variance of X denoted by $V(X)$ is given by

$$V(X) = E[X - E(X)]^2 = E[X - \mu]^2$$
$$= E[X^2 + [E(X)]^2 - 2\,XE(X)]$$
$$V(X) = E(X^2) + [E(X)]^2 - 2\,E(X)E(X)$$
$$= E(X^2) + [E(X)]^2 - 2[E(X)]^2$$
$$\therefore V(X) = E(X^2) - [E(X)]^2$$

Properties of Variance

1. If X is a random variable then $V(aX + b) = a^2V(X)$, where a and b are constants.

(JNTU 2001, 2007 Nov)

Proof: Let $Y = aX + b$

$$E(Y) = E(aX + b) = aE(X) + b$$
$$Y - E(Y) = a[X - E(X)]$$

Squaring on both the sides we get,

$$[Y - E(Y)]^2 = a^2[X - E(X)]^2$$

Applying expectation on both the sides we get,

$$E[Y - E(Y)]^2 = a^2E[X - E(X)]^2$$
$$V(Y) = a^2V(X)$$
$$V(aX + b) = a^2V(X)$$

2. If $b = 0$ and $a = 1$ the above relation becomes,
   $V(X) = V(X)$
   If $b = 1$ then $V(aX + 1) = a^2V(X) + 0$
   $a = 1$, $V(X + b) = V(X)$
   Hence, variance is independent of change of scale but not change of origin.
   Caution:
   - The positive square root of the variance, σ is called the standard deviation of X.
   - The quantity $x - \mu$ is called the deviation of an observation from its mean.

Worked Out Examples

EXAMPLE 3.3

The random variable X representing the number of errors per 100 lines of software code has the following probability distribution:

x	2	3	4	5	6
$f(x)$	0.01	0.25	0.4	0.3	0.04

Solution: The formula used for calculation of variance is

$$V(X) = E(X^2) - [E(X)]^2$$

$$E(X^2) = \sum_{x=2}^{6} x^2 f(x)$$

$$E(X^2) = 2^2(0.01) + 3^2(0.25) + 4^2(0.4) + 5^2(0.3) + 6^2(0.04)$$
$$= 0.04 + 2.25 + 6.4 + 7.5 + 1.44$$
$$= 17.63$$

$$E(X) = \sum_{x=2}^{6} xf(x)$$
$$= 2(0.01) + 3(0.25) + 4(0.4) + 5(0.3) + 6(0.04)$$
$$= 0.02 + 0.75 + 1.6 + 1.5 + 0.24$$
$$= 4.11$$

$$\therefore V(X) = E(X^2) - [E(X)]^2 = 17.63 - (4.11)^2$$
$$= 17.63 - 16.8921$$
$$= 0.7379$$

EXAMPLE 3.4

The probability function of a random variable X is as follows:

x	1	2	3	4	5
$f(x)$	0.02	0.1	0.6	0.2	0.08

(i) Determine mean of X.

(ii) Determine variance of X.

Solution:

(i) $$E(X) = \sum_{x=1}^{5} xf(x) = 1(0.02) + 2(0.1) + 3(0.6) + 4(0.2) + 5(0.08)$$
$$= 0.02 + 0.2 + 1.8 + 0.8 + 0.4$$
$$= 3.22$$

(ii) $$E(X^2) = \sum_{x=1}^{5} x^2 f(x) = 1(0.02) + 2^2(0.1) + 3^2(0.6) + 4^2(0.2) + 5^2(0.08)$$
$$= 0.02 + 0.4 + 5.4 + 3.2 + 2$$
$$= 11.02$$

$$\therefore V(X) = E(X) - [E(X)]^2 = 11.02 - (3.22)^2$$
$$= 11.02 - 10.3684$$
$$= 0.6516$$

Worked Out Examples

EXAMPLE 3.5

In a lottery, there are 200 prizes of ₹5, 20 prizes of ₹25, and 5 prizes of ₹100. Assuming that 10,000 tickets are to be issued and sold, what is a fair price to pay for a ticket?

Solution: The probabilities are obtained for each of the prizes.

$$P(X=5)=\frac{200}{10000}=0.02,\ P(X=25)=\frac{20}{10000}=0.002$$

$$P(X=100)=\frac{5}{10000}=0.0005$$

Since the sum of probabilities should be 1, the last one can be taken as

$$\begin{aligned}P(X=0) &= 1-(0.02+0.002+0.0005)\\ &= 1-0.0225\\ &= 0.9775\end{aligned}$$

Hence, $P(X=0)=0.9775$

$\therefore$ The probability distribution of X is given by,

X	5	25	100	0
$P(X=x)$	0.02	0.002	0.0005	0.9775

$$E(X)=\Sigma xP(X=x)=5(0.02)+25(0.002)+100(0.0005)+0$$
$$=0.1+0.05+0.005=0.2$$
$$E(X^2)=\Sigma x^2P(X=x)=5^2(0.02)+25^2(0.002)+100^2(0.0005)$$
$$=0.5+1.25+0.05=1.8$$
$$V(X)=E(X^2)-[E(X)]^2=1.8-(0.2)^2=1.76$$

Work Book Exercises

1. Let X be a random variable with the following distribution:

x	−1	0	1	2
$P(X=x)$	0.4	0.3	0.2	0.1

   (i) Determine the mean of X.
   (ii) Determine the variance of X.

2. A random variable X has the following probability function:

Values of X, x	−2	−1	0	1	2	3
$P(X=x)$	0.1	K	0.2	$2K$	0.3	K

   (i) Find the value of K and calculate the mean and variance.
   (ii) Construct the cdf $F(x)$. (JNTU 2008 April Set 3)

3. Calculate the expectation and variance of X. If the probability distribution of the random variable X is given by:

X	−1	0	1	2	3
f	0.3	0.1	0.1	0.3	0.2

(JNTU 2006)

4. Let $Y = X^2 + 2X$. The probability for X is given as follows:

x	0	1	2	3
$P(x)$	0.1	0.3	0.5	0.1

(i) Find the probability function of Y.

(ii) Find mean and variance of Y.

[**Ans.:** 6.4, 16.24]

5. (i) Find the expectation of the number on a die when thrown.

(ii) Two unbiased dice are thrown. Find the expected values of the sum of number of points on them.

Worked Out Example

EXAMPLE 3.6

The total number of hours, measured in units of 100 hours that a family runs a vacuum cleaner over a period of one year is a continuous random variable X that has the density function.

$$f(x) = \begin{cases} x, & 0 < x < 1 \\ 2 - x, & 1 < x < 2 \\ 0, & \text{elsewhere} \end{cases}$$

(i) Find the average number of hours per year that families run their vacuum cleaners

(ii) Find the variance.

Solution: Let X denote the hours per year that families run their vacuum cleaners.

(i) Average number of hours per year

$$E(X) = \int_{-\infty}^{\infty} x f(x)\, dx$$

$$= \int_0^1 x f(x)\, dx + \int_1^2 x f(x)\, dx$$

$$= \int_0^1 x.x\, dx + \int_1^2 x(2-x)\, dx = \int_0^1 x^2\, dx + \int_1^2 (2x - x^2)\, dx$$

$$E(X) = \left|\frac{x^3}{3}\right|_0^1 + 2\left|\frac{x^2}{2}\right|_1^2 - \left|\frac{x^3}{3}\right|_1^2$$

$$= \frac{1}{3} + (4-1) - \left(\frac{8}{3} - \frac{1}{3}\right) = 1$$

$$E(X^2) = \int_{-\infty}^{\infty} x^2 f(x)\, dx$$

$$= \int_0^1 x^2 . x\ dx + \int_1^2 x^2 (2-x)\ dx = \int_0^1 x^3\ dx + \int_1^2 (2x^2 - x^3)\ dx$$

$$= \left|\frac{x^4}{4}\right|_0^1 + \left|2\frac{x^3}{3} - \frac{x^4}{4}\right|_1^2 = \frac{1}{4} + \frac{2}{3}(8-1) - \frac{1}{4}(16-1)$$

$$= \frac{1}{4} + \frac{14}{3} - \frac{15}{4} = 0.25 + 4.666 - 3.75 = 1.166$$

(ii) Variance of $X = V(X) = E(X^2) - [E(X)]^2$

$= 1.166 - 1^2$

$= 0.166$

So far we have seen expectations of a random variable. Now let us learn about the expectations of their functions.

3.3 EXPECTATION OF A FUNCTION OF RANDOM VARIABLES

Let X be a discrete random variable with probability function $f(x)$. Then $Y = g(X)$ is also a discrete random variable and the probability function of Y is

$$h(y) = P(Y = y) = \sum_{\{x|g(x)=y\}} P(X = x) = \sum_{\{x|g(x)=y\}} f(x)$$

If X takes the values $x_1, x_2, \ldots x_n$ and y takes the values $y_1, y_2, \ldots y_m (m \le n)$, then

$$E\left[g(X)\right] = \sum_{i=1}^{n} g(x_i) f(x_i)$$

Similarly, if X is a continuous random variable having probability density $f(x)$, then

$$E[g(X)] = \int_{-\infty}^{\infty} g(x)\ f(x)\, dx$$

Suppose, $f(x, y)$ is joint density function of two continuous random variables, then expectation of $g(x, y)$ is given by $E[g(X,Y)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y)\ f(x,y)\, dxdy$

$= \sum_x \sum_y g(x,y) f(x,y)$ if x, y are discrete random variables.

3.4 VARIANCE FOR JOINT DISTRIBUTIONS

Suppose X and Y are two dimensional random variables, then

$$E(X) = \sum_x \sum_y x f(x,y) = \sum_x x g(x) \quad \text{(Discrete case)}$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x f(x,y)\, dxdy = \int_{-\infty}^{\infty} x g(x)\, dx$$

$$E(X^2) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x^2 f(x,y)\, dxdy, V(X) = E(X^2)[E(X)]^2 \quad \text{(Continuous case)}$$

$$E(Y) = \sum_x \sum_y y f(x,y) = \sum_y y h(y) \quad \text{(Discrete case)}$$

$$V(X) = E(X^2) - [E(X)]^2 \quad \text{(Continue case)}$$

$$E(Y) = \sum_x \sum_y y f(x,y) = \sum_y y h(y) \quad \text{(Discrete case)}$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y f(x,y)\, dxdy = \int_{-\infty}^{\infty} y h(y)\, dy \quad \text{(Continuous case)}$$

where $g(x)$ and $h(y)$ are marginal distributions of X and Y, respectively.

Worked Out Examples

EXAMPLE 3.7

The proportion of people who respond to a certain mail-order solicitation is a continuous random variable X that has the density function,

$$f(x) = \begin{cases} \dfrac{2(x+2)}{5}, & 0 < x < 1 \\ 0, & \text{elsewhere} \end{cases}$$

(i) Find the expected value of X.

(ii) Find the variance of X.

Solution:

$E(X)$ = Expected value of X

$$= \int_{-\infty}^{\infty} x f(x)\, dx = \int_0^1 x \frac{2}{5}(x+2)\, dx$$

$$= \frac{2}{5}\int_0^1 x(x+2)\, dx = \frac{2}{5}\left[\frac{x^3}{3} + \cancel{2}\frac{x^2}{\cancel{2}}\right]_0^1$$

$$= \frac{2}{5}\left[\frac{1}{3} + 1\right] = \frac{4}{15}$$

$$E(X^2) = \int_0^1 x^2 f(x)\,dx = \int_0^1 x^2 \frac{2}{5}(x+2)\,dx$$

$$= \frac{2}{5}\int_0^1 (x^3 + 2x^2)\,dx$$

$$= \frac{2}{5}\left[\frac{x^3}{4} + 2\frac{x^3}{3}\right]_0^1$$

$$= \frac{2}{5}\left[\frac{1}{4} + \frac{2}{3}\right]$$

$$= \frac{2}{5}\left(\frac{3+8}{12}\right) = \frac{22}{60} = \frac{11}{30}$$

Variance of $X = V(X) = E(X^2) - [E(X)]^2$

$$= \frac{11}{30} - \left(\frac{4}{15}\right)^2$$

$$= 0.3666 - 0.0711$$

$$V(X) = 0.295$$

EXAMPLE 3.8

Let X and Y be random variables with the following joint distribution:

X \ Y	−3	2	4	Total
1	0.1	0.2	0.2	0.5
3	0.3	0.1	0.1	0.5
Total	0.4	0.3	0.3	1

(i) Find the marginal distributions of X and Y.

(ii) Find $E(X)$, $E(Y)$, $E(XY)$.

Solution:

(i) The marginal distribution of X:

X	1	3
$g(X)$	0.5	0.5

The marginal distribution of Y:

Y	−3	2	4
$h(Y)$	0.4	0.3	0.3

(ii) $E(X) = \Sigma x g(x)$

$= 1(0.5) + 3(0.5) = 2$

$E(Y) = \Sigma y h(y)$

$= -3(0.4) + 2(0.3) + 4(0.3)$

$= 0.6$

$E(XY) = \sum_x \sum_y xyf(x,y)$

$= 1(-3)\,(0.1) + 1(2)\,(0.2) + 1(4)\,(0.2) + 3(-3)\,(0.3) + 3(2)\,(0.1) + 3(0.1)\,(4) = 0$

EXAMPLE 3.9

Consider the following probability distribution:

X \ Y	0	1	2
0	0.1	0.2	0.1
1	0.2	0.3	0.1

Determine $E(X)$, $E(Y)$, $V(X)$, $E(4X + 5)$, and $V(2X + 3)$

Solution: The marginal distribution of X is

X	0	1
$g(x) = P(X = x)$	0.1 + 0.2 + 0.1 = 0.4	0.2 + 0.3 + 0.1 = 0.6

That is, the marginal distribution of X is

X	0	1
$g(x)$	0.4	0.6

The marginal distribution of Y is

Y	0	1	2
$h(y)$	0.1 + 0.2 = 0.3	0.2 + 0.3 = 0.5	0.1 + 0.1 = 0.2

That is, the marginal distribution of Y is

Y	0	1	2
$h(y)$	0.3	0.5	0.2

$\therefore E(X) = \Sigma x g(x) = 0(0.4) + 1(0.6) = 0.6$

$E(X^2) = \Sigma x^2 g(x) = 0^2(0.4) + 1^2(0.6) = 0.6$

$\therefore V(X) = E(X^2) - [E(X)]^2 = 0.6 - (0.6)^2 = 0.24$

$E(Y) = \Sigma y h(y) = 0(0.3) + 1(0.5) + 2(0.2) = 0.9$

$$E(Y^2) = \Sigma y^2 h(y) = 0(0.3) + 1^2(0.5) + 2^2(0.2) = 1.3$$
$$\therefore V(Y) = E(Y^2) - [E(Y)]^2 = 1.3 - (0.9)^2 = 0.49$$

From the properties of expectation and variance, we have

$$\begin{aligned} E(4X + 5) &= 4E(X) + 5 \\ &= 4(0.6) + 5 \\ &= 7.4 \\ V(2X + 3) &= 4V(X) \\ &= 4(0.24) \\ &= 0.96 \end{aligned}$$

Hence, $E(X) = 0.6$, $E(Y) = 0.9$, $V(X) = 0.24$, $E(4X + 5) = 7.4$, and $V(2X + 3) = 0.96$.

3.5 COVARIANCE

If X and Y are two random variables, then covariance between them is given by,

$$\begin{aligned} \text{Cov}(X, Y) &= E\{[X - E(X)][Y - E(Y)]\} \\ &= E\{XY - XE(Y) - YE(X) + E(X)E(Y)\} \\ &= E(XY) - E(X)E(Y) - E(Y)E(X) + E(X)E(Y) \\ &= E(XY) - E(X)E(Y) \end{aligned}$$

If X and Y are independent random variables, then

$$\begin{aligned} \text{Cov}(X, Y) &= E(X)\,E(Y) - E(X)\,E(Y) \\ &= 0 \ \{\text{Since if } X \text{ and } Y \text{ are independent } E(XY) = E(X)E(Y)\} \end{aligned}$$

Results

RESULT 1: $\text{Cov}(aX, bY) = ab\text{Cov}(X, Y)$.

Proof: $\begin{aligned} \text{Cov}(aX, bY) &= E(aX \cdot bY) - E(aX) \cdot b(Y) \\ &= abE(XY) - a\,E(X) \cdot bE(Y) \\ &= ab[E(XY) - E(X)E(Y)] \end{aligned}$

$$\text{Cov}(aX, bY) = ab\text{Cov}(X, Y)$$

RESULT 2: $\text{Cov}(X + a, Y + b) = \text{Cov}(X, Y)$

Proof: $\begin{aligned} \text{Cov}(X + a, Y + b) &= E[(X + a)(Y + b)] - E[(X + a)E(Y + b)] \\ &= E[XY + bX + aY + ab] - [E(X) + a]\,[E(Y) + b] \\ &= E(XY) + bE(X) + aE(Y) + ab - E(X)E(Y) - bE(X) - aE(Y) - ab \end{aligned}$

$$\text{Cov}(X + a, y + b) = E(XY) - E(X)E(Y)$$

RESULT 3: $\text{Cov}(aX + b, cY + d) = ac\text{Cov}(X, Y)$

Proof: $\begin{aligned} \text{Cov}(aX + b, cY + d) &= E[(aX + b)\,(cY + d)] - E(aX + b)E(cY + d) \\ &= E[acXY + adX + bcY + bd] - [E(aX) + b][E(cY) + d] \\ &= acE(XY) + adE(X) + bcE(Y) + bd - acE(X)E(Y) - adE(X) - bcE(Y) - bd \\ &= acE(XY) - acE(X)E(Y) \end{aligned}$

$$\therefore \text{Cov}(aX + b, cY + \text{d}) = ac[E(XY) - E(X)E(Y)]$$
$$\text{Cov}(aX + b, cY + \text{d}) = ac\text{Cov}(X, Y)$$

Cauchy-Schwartz Inequality

For any two random variables X and Y,

$$|E(XY)|^2 \le E(X^2)E(Y^2)$$

3.6 CONDITIONAL EXPECTATION

If X and Y have joint density function $f(x, y)$, then the conditional density function of Y given X is $f\left(\frac{y}{x}\right) = \frac{f(x,y)}{g(x)}$ where $g(x)$ is the marginal density of X.

The conditional expectation or conditional mean of Y given X is

$$E\left(\frac{Y}{X} = x\right) = \int_{-\infty}^{\infty} yf\left(\frac{y}{x}\right) dy \qquad \text{(Continuous case)}$$

$$= \sum_{y} yf\left(\frac{y}{x}\right) \qquad \text{(Discrete case)}$$

If X and Y are independent random variables, then

$$E\left(\frac{Y}{X} = x\right) = E(Y) \text{ and } E(Y) = \int_{-\infty}^{\infty} E\left(\frac{Y}{X} = x\right) g(x)\, dx$$

Worked Out Examples

EXAMPLE 3.10

The joint probability function for the random variables X and Y is given as follows:

X \ Y	0	1	2
0	$\frac{1}{18}$	$\frac{1}{9}$	$\frac{1}{6}$
1	$\frac{1}{9}$	$\frac{1}{18}$	$\frac{1}{9}$
2	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{18}$

(i) Find the marginal densities of X and Y.

(ii) Find the conditional expectation of Y given X.

(iii) Find the conditional expectation of X given Y.

Solution:

(i) The marginal densities of X are

$$g(X = 0) = \frac{1}{18} + \frac{1}{9} + \frac{1}{6} = \frac{1+2+3}{18} = \frac{6}{18} = \frac{1}{3}$$

$$g(X = 1) = \frac{1}{9} + \frac{1}{18} + \frac{1}{9} = \frac{2+1+2}{18} = \frac{5}{18}$$

$$g(X=2)=\frac{1}{6}+\frac{1}{6}+\frac{1}{18}=\frac{3+3+1}{18}=\frac{7}{18}$$

The marginal distribution of X is

X	0	1	2
$g(x)=P(X=x)$	$\frac{6}{18}$	$\frac{5}{18}$	$\frac{7}{18}$

The marginal densities of Y are

$$h(y=0)=\frac{1}{18}+\frac{1}{9}+\frac{1}{6}=\frac{1+2+3}{18}=\frac{6}{18}=\frac{1}{3}$$

$$h(y=1)=\frac{1}{9}+\frac{1}{18}+\frac{1}{6}=\frac{2+1+3}{18}=\frac{6}{18}=\frac{1}{3}$$

$$h(y=2)=\frac{1}{6}+\frac{1}{9}+\frac{1}{18}=\frac{3+2+1}{18}=\frac{6}{18}=\frac{1}{3}$$

The marginal distribution of Y is

Y	0	1	2
$h(y)=P(Y=y)$	$\frac{6}{18}$	$\frac{6}{18}$	$\frac{6}{18}$

(ii) The conditional expectation of Y given X

$$\mathrm{E}\left(\frac{Y}{X}\right)=y\,f\left(\frac{y}{x}\right)$$

First lets us find the conditional densities of $\frac{Y}{X}$.

$$f\left(\frac{Y=0}{X=0}\right)=\frac{f(X=0,Y=0)}{g(X=0)}=\frac{\frac{1}{18}}{\frac{6}{18}}=\frac{1}{6}$$

$$f\left(\frac{Y=1}{X=0}\right)=\frac{f(X=0,Y=1)}{g(X=0)}=\frac{\frac{1}{9}}{\frac{6}{18}}=\frac{2}{6}$$

$$f\left(\frac{Y=2}{X=0}\right)=\frac{f(X=0,Y=2)}{g(X=0)}=\frac{\frac{1}{6}}{\frac{6}{18}}=\frac{3}{6}$$

Conditional distribution of Y given $X=0$

Y	0	1	2
$f\left(\frac{Y}{X}=0\right)$	$\frac{1}{6}$	$\frac{2}{6}$	$\frac{3}{6}$

$$\mathrm{E}\left(\frac{Y}{X}=0\right)=\sum yf\left(\frac{Y}{X}=0\right)$$
$$=0\left(\frac{1}{6}\right)+1\left(\frac{2}{6}\right)+2\left(\frac{3}{6}\right)$$
$$=\frac{2}{6}+1=\frac{8}{6}=\frac{4}{3}$$

$$f\left(\frac{Y=0}{X=1}\right)=\frac{f(X=1,\ Y=0)}{g(X=1)}=\frac{\frac{1}{9}}{\frac{5}{18}}=\frac{2}{5}$$

$$f\left(\frac{Y=1}{X=1}\right)=\frac{f(X=1,Y=1)}{g(X=1)}=\frac{\frac{1}{18}}{\frac{5}{18}}=\frac{1}{5}$$

$$f\left(\frac{Y=2}{X=1}\right)=\frac{f(X=1,Y=2)}{g(X=1)}=\frac{\frac{1}{9}}{\frac{5}{18}}=\frac{2}{5}$$

Conditional distribution of Y given $X=1$

Y	0	1	2
$f\left(\frac{Y}{X}=x\right)$	$\frac{2}{5}$	$\frac{1}{5}$	$\frac{2}{5}$

Conditional expectation of Y given $X=1$ is

$$E\left(\frac{Y}{X}=1\right)=\sum yf\left(\frac{Y}{X}=1\right)$$
$$=0\left(\frac{2}{5}\right)\ +\ 1\left(\frac{1}{5}\right)\ +\ 2\left(\frac{2}{5}\right)$$
$$=\frac{1}{5}\ +\ \frac{4}{5}=1$$

$$f\left(\frac{Y=0}{X=2}\right)=\frac{f(X=2,Y=0)}{g(X=2)}=\frac{\frac{1}{6}}{\frac{7}{18}}=\frac{3}{7}$$

$$f\left(\frac{Y=1}{X=2}\right)=\frac{f(X=2,Y=1)}{g(X=2)}=\frac{\frac{1}{6}}{\frac{7}{18}}=\frac{3}{7}$$

$$f\left(\frac{y=2}{x=2}\right)=\frac{f(X=2,Y=2)}{g(X=2)}=\frac{\frac{1}{18}}{\frac{7}{18}}=\frac{1}{7}$$

Conditional distribution of Y given $X = 2$

Y	0	1	2
$f\left(\frac{Y}{X}=2\right)$	$\frac{3}{7}$	$\frac{3}{7}$	$\frac{1}{7}$

Conditional expectation of Y given $X = 2$ is

$$E\left(\frac{Y}{X}=2\right)=\sum yf\left(\frac{Y}{X}=2\right)$$

$$=0\left(\frac{3}{7}\right)+1\left(\frac{3}{7}\right)+2\left(\frac{1}{7}\right)$$

$$=\frac{3}{7}+\frac{2}{7}=\frac{5}{7}$$

Hence, the conditional expectation of Y given X is

X	0	1	2
$E\left(\frac{Y}{X}=2\right)$	$\frac{4}{3}$	1	$\frac{5}{7}$

(iii) The conditional expectation of X given Y is

$$E\left(\frac{X}{Y}\right)=xf\left(\frac{X}{Y}\right)$$

First we have to find the conditional densities of $\frac{X}{Y}$.

$$f\left(\frac{X=0}{Y=0}\right)=\frac{f(X=0,Y=0)}{h(Y=0)}=\frac{\frac{1}{18}}{\frac{6}{18}}=\frac{1}{6}$$

$$f\left(\frac{X=1}{Y=0}\right)=\frac{f(X=1,Y=0)}{h(Y=0)}=\frac{\frac{1}{9}}{\frac{6}{18}}=\frac{2}{6}$$

$$f\left(\frac{X=2}{Y=0}\right)=\frac{f(X=2,Y=0)}{h(Y=0)}=\frac{\frac{1}{6}}{\frac{6}{18}}=\frac{3}{6}$$

Conditional distribution of X given $Y = 0$ is

X	0	1	2
$f\left(\frac{X}{Y}\right)=0$	$\frac{1}{6}$	$\frac{2}{6}$	$\frac{3}{6}$

Conditional expectation of X given $Y = 0$ is

$$E\left(\frac{X}{Y}=0\right)=\sum xf\left(\frac{X}{Y}\right)$$

$$=0\left(\frac{1}{6}\right)+1\left(\frac{2}{6}\right)+2\left(\frac{3}{6}\right)$$

$$=\frac{2}{6}+\frac{6}{6}=\frac{4}{3}$$

$$f\left(\frac{X=0}{Y=1}\right)=\frac{f(X=0,Y=1)}{h(Y=1)}=\frac{\frac{1}{9}}{\frac{6}{18}}=\frac{2}{6}$$

$$f\left(\frac{X=1}{Y=1}\right)=\frac{f(X=1,Y=1)}{h(Y=1)}=\frac{\frac{1}{18}}{\frac{6}{18}}=\frac{1}{6}$$

$$f\left(\frac{X=2}{Y=1}\right)=\frac{f(X=2,Y=1)}{h(Y=1)}=\frac{\frac{1}{6}}{\frac{6}{18}}=\frac{3}{6}$$

The conditional distribution of X given $Y = 1$

X	0	1	2
$E\left(\frac{X}{Y}=1\right)$	$\frac{2}{6}$	$\frac{1}{6}$	$\frac{3}{6}$

The conditional expectation of X given $Y = 1$

$$E\left(\frac{X}{Y}=1\right)=\sum xf\left(\frac{X}{Y}=1\right)$$

$$=0\left(\frac{2}{6}\right)+1\left(\frac{1}{16}\right)+2\left(\frac{3}{6}\right)$$

$$=\frac{1}{16}+\frac{6}{6}=\frac{7}{6}$$

$$f\left(\frac{X=0}{Y=2}\right)=\frac{f(X=0,Y=2)}{h(Y=2)}=\frac{\frac{1}{6}}{\frac{6}{18}}=\frac{3}{6}$$

$$f\left(\frac{X=1}{Y=2}\right)=\frac{f(X=1,Y=2)}{h(Y=2)}=\frac{\frac{1}{9}}{\frac{6}{18}}=\frac{2}{6}$$

$$f\left(\frac{X=2}{Y=2}\right)=\frac{f(X=2,Y=2)}{h(Y=2)}=\frac{\frac{1}{18}}{\frac{6}{18}}=\frac{1}{6}$$

The conditional distribution of X given $Y=2$

X	0	1	2
$f\left(\frac{X}{Y}=2\right)$	$\frac{3}{6}$	$\frac{2}{6}$	$\frac{1}{6}$

The conditional expectation of X given $Y=2$ is

$$E\left(\frac{X}{Y}=2\right)=\sum xf\left(\frac{X}{Y}=2\right)$$
$$=0\left(\frac{3}{6}\right)+1\left(\frac{2}{6}\right)+2\left(\frac{1}{6}\right)$$
$$=\frac{2}{6}+\frac{2}{6}=\frac{4}{6}=\frac{2}{3}$$

Hence, the conditional expectation of X given $Y=2$ is

X	0	1	2
$E\left(\frac{X}{Y}\right)$	$\frac{4}{3}$	$\frac{7}{6}$	$\frac{2}{3}$

EXAMPLE 3.11

Let X and Y have joint density function,

$$f(x,y)=\begin{cases}x+y, & 0\le x\le 1, \quad 0\le y\le 1\\ 0, & \text{otherwise}\end{cases}$$

(i) Find the marginal densities of X and Y.

(ii) Find the conditional expectation of X given Y.

(iii) Find the conditional expectation of Y given X.

Solution:

(i) The marginal density of X is

$$g(x)=\int_{-\infty}^{\infty} f(x,y)\,dy=\int_{0}^{1}(x+y)\,dy$$
$$=x(y)_0^1+\frac{y^2}{2}\Big|_0^1=x+\frac{1}{2}$$
$$g(x)=\begin{cases}x+\frac{1}{2}, & 0\le x\le 1\\ 0, & \text{otherwise}\end{cases}$$

The marginal density of Y is,

$$h(y) = \int_{-\infty}^{\infty} f(x,y)\,dx = \int_{0}^{1} (x+y)\,dx$$

$$= \frac{x^2}{2}\Big|_0^1 + y(x)\Big|_0^1 = \frac{1}{2} + y$$

$$\therefore h(y) = \begin{cases} y + \frac{1}{2}, & 0 \le y \le 1 \\ 0, & \text{otherwise} \end{cases}$$

The conditional density of X given Y is

$$f\left(\frac{X}{Y}\right) = \frac{f(x,y)}{h(y)} = \frac{x+y}{y+\frac{1}{2}}$$

The conditional density of Y given X is

$$f\left(\frac{Y}{X}\right) = \frac{f(x,y)}{g(x)} = \frac{x+y}{x+\frac{1}{2}}$$

(ii) The conditional expectation of X given Y is

$$E\left(\frac{X}{Y}\right) = \int_{-\infty}^{\infty} x f\left(\frac{x}{y}\right) dx = \int_{0}^{1} \frac{x(x+y)}{y+\frac{1}{2}}\,dx$$

$$= \frac{1}{y+\frac{1}{2}} \int_0^1 (x^2 + xy)\,dx = \frac{1}{y+\frac{1}{2}} \left| \frac{x^3}{3} + y\,\frac{x^2}{2} \right|_0^1$$

$$= \frac{\frac{1}{3} + \frac{y}{2}}{y+\frac{1}{2}} = \frac{\frac{2+3y}{6}}{\frac{2y+1}{2}} = \frac{2+3y}{3(2y+1)}$$

$$\therefore E\left(\frac{X}{Y}\right) = \frac{2+3y}{6y+3}, \quad 0 \le y \le 1$$

(iii) The conditional expectation of Y given X is

$$E\left(\frac{Y}{X}\right) = \int_{-\infty}^{\infty} y f\left(\frac{y}{x}\right) dy = \int_{0}^{1} \frac{y(x+y)}{x+\frac{1}{2}}\,dy$$

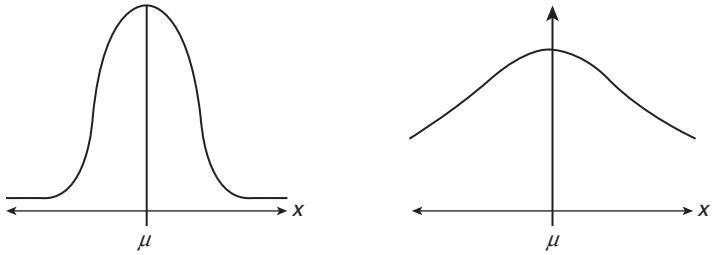
$$= \frac{1}{x+\frac{1}{2}} \int_0^1 (xy + y^2)\,dy = \frac{1}{x+\frac{1}{2}} \left[x \left| \frac{y^2}{2} \right|_0^1 + \left| \frac{y^3}{3} \right|_0^1 \right]$$

$$E\left(\frac{Y}{X}\right) = \frac{\frac{x}{2}+\frac{1}{3}}{x+\frac{1}{2}} = \frac{3x+2}{3(2x+1)}$$

$$\therefore E\left(\frac{Y}{X}\right) = \begin{cases} \frac{3x+2}{6x+3}, & 0 \le x \le 1 \\ 0, & \text{otherwise} \end{cases}$$

3.7 CHEBYCHEV'S INEQUALITY

In the earlier sections, we have seen that variance tells about the variability of the observations about the mean. If a random variable has a smaller variance or standard deviation, we can expect most of the values centred around mean. Hence, the probability that a random variable assumes a value within a certain interval about the mean is greater than for a similar random variable with a larger standard deviation.



(i)(a) (i)(b)

Figure (i) shows that variability of observations about the mean when the standard deviation is small, most of the area (probability of observation) lies close to μ as in Fig. (i)(a), whereas if the standard deviation is large, the area is spread not closer to μ as in Fig. (i)(b).

Hence, the fraction of the area between any two values symmetric about the mean is related to the standard deviations.

The following theorem gives an estimate of the probability that a random variable assumes a value within K standard derivations of its mean for any real number K.

Chebychev's Theorem

If X is a random variable with mean μ and variance σ^2, then for any positive number K, we have

$$P\{|X-\mu| < K\sigma\} \ge 1-\frac{1}{K^2} \quad \text{or}$$

$$P\{\mu - K\sigma < X < \mu + K\sigma\} \ge 1-\frac{1}{K^2} \quad \text{or}$$

$$P\{|X-\mu| \ge K\sigma\} \le \frac{1}{K^2}$$

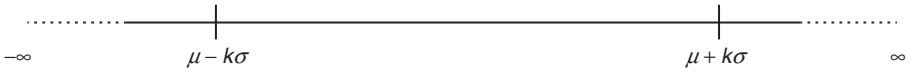
In other words, the probability that any random variable X will assume a value within K standard deviations of the mean is at least $\left(1-\frac{1}{K^2}\right)$.

Proof: Let X be continuous random variable. The variance of a random variable X is given by

$$\sigma^2 = \sigma_X^2 = E[X - E(X)]^2 = E(X - \mu)^2$$

$$= \int_{-\infty}^{\infty} (x - \mu)^2 f(x)\, dx$$

where $f(x)$ is the probability density function of the random variable X. Let us split the limits of integration as follows:



$$\sigma^2 = \int_{-\infty}^{\mu-K\sigma} (x - \mu)^2 f(x)\, dx + \int_{\mu-K\sigma}^{\mu+K\sigma} (x - \mu)^2 f(x)\, dx + \int_{\mu+K\sigma}^{\infty} (x - \mu)^2 f(x)\, dx$$

Since R.H.S is sum of these positive terms L.H.S should be greater than or equal to two terms on the R.H.S.

$$\therefore \sigma^2 \geq \int_{-\infty}^{\mu-K\sigma} (x - \mu)^2 f(x)\, dx + \int_{\mu+K\sigma}^{\infty} (x - \mu)^2 f(x)\, dx$$

From the above two integers, $x \leq \mu - K\sigma$

$$x \geq \mu + K\sigma$$

which implies $|x - \mu| \geq K\sigma$

Substituting in the integrals, we get

$$\sigma^2 \geq K^2\sigma^2 \left[\int_{-\infty}^{\mu-K\sigma} f(x)\, dx + \int_{\mu+K\sigma}^{\infty} f(x)\, dx \right]$$

$$= K^2\sigma^2 [P(X \leq \mu - K\sigma) + P(X \geq \mu + K\sigma)]$$

$$\sigma^2 = K^2\sigma^2 \left[P(|X - \mu| \geq K\sigma) \right]$$

$$\frac{1}{K^2} \geq P\left[|X - \mu| \geq K\sigma \right]$$

$$\therefore P\left[|X - \mu| \geq K\sigma \right] \leq \frac{1}{K^2}$$

Also, since

$$P\left[|X - \mu| \geq K\sigma \right] + P\left[|X - \mu| < K\sigma \right] = 1$$

$$\therefore P\left[|X - \mu| < K\sigma \right] = 1 - P\left[|X - \mu| \geq K\sigma \right]$$

$$\geq 1 - \frac{1}{K^2}$$

Caution:

- If we take $K\sigma = C > 0$, the above inequalities reduce to $P\{|X-\mu| \geq C\} \leq \frac{\sigma^2}{C^2}$ and $P\{|X-\mu| < C\} \geq 1-\frac{\sigma^2}{C^2}$

$$\Rightarrow P\{|X-E(X)| \geq C\} \leq \frac{\text{Var}(X)}{C^2} \text{ and } \Rightarrow P\{|X-E(X)| < C\} \geq 1-\frac{\text{Var}(X)}{C^2}$$

- This theorem holds for any distribution of observations. The value given by the theorem is a lower bound only.

Worked Out Examples

EXAMPLE 3.12

A random variable X has a mean $\mu = 10$ and a variance $\sigma^2 = 4$. Using Chebychev's theorem, find

(i) $P(|X-10| \geq 3)$

(ii) $P(|X-10| < 3)$

(iii) $P(5 < X < 15)$

(iv) The value of the constant C such that $P(|X-10| \geq C) \leq 0.04$

Solution:

(i) Chebychev's theorem is given by

$$P(|X-\mu| \geq K\sigma) \leq 1/K^2$$

Given $\mu = 10$, $\sigma^2 = 4$

$$K\sigma = K(2) = 3 \Rightarrow K = \frac{3}{2}$$

$$\therefore P(|X-10| \geq 3) \leq \frac{1}{\left(\frac{3}{2}\right)^2} = \frac{4}{9}$$

(ii) Another form of Chebychev's theorem is

$$P(|X-\mu| < K\sigma) \geq 1-\frac{1}{K^2}$$

$$P(|X-10| < 3) \geq 1-\frac{4}{9} = \frac{5}{9}$$

(iii) $P(\mu - K\sigma < X < \mu + K\sigma) \geq 1-\frac{1}{K^2}$

$$P(5 < X < 15) \geq 1-\frac{1}{K^2}$$

$$10 - 2(K) = 5 \qquad \mu + K\sigma = 15$$
$$5 = 2K \Rightarrow K = 2.5 \qquad 10 + K(2) = 15$$
$$K = 2.5$$

$$\therefore P(5 < X < 15) \geq 1 - \frac{1}{\left(\frac{5}{2}\right)^2} = 1 - \frac{4}{25} = \frac{21}{25}$$

(iv) To find C given that

$$P(|1 \times -10| \geq C) \leq 0.04 \qquad \frac{1}{K^2} = 0.04$$

$$K\sigma = C$$
$$5(2) = C \qquad K^2 = \frac{1}{0.04} = 25$$
$$\therefore\ C = 10 \qquad K = 5$$

EXAMPLE 3.13

Suppose X is a normal variate with mean μ and standard distribution σ. Determine the Chebychev's theorem for $K = 2$ and $K = 3$.

Solution: Chebychev's theorem is

$$P(\mu - K\sigma < X < \mu + K\sigma) = \int_{\mu - K\sigma}^{\mu + K\sigma} f(x)\, dx \geq 1 - \frac{1}{K^2}$$

For $K = 2$,

$$P(\mu - 2\sigma < X < \mu + 2\sigma) \geq 1 - \frac{1}{4} = \frac{3}{4}$$

For a normal variate

$$P(\mu - 2\sigma < X < \mu + 2\sigma) = 0.9544 \text{ which is greater than } 0.75.$$

For $K = 3$,

$$P(\mu - 3\sigma < X < \mu + 3\sigma) \geq 1 - \frac{1}{9} = \frac{8}{9} = 0.888$$

For a normal variate,

$$P(\mu - 3\sigma < X < \mu + 3\sigma) = 0.9974 \text{ which is greater than } 0.88.$$

3.8 MOMENTS

Central Moments

The r^{th} moment of a random variable X about the mean μ is called r^{th} central moment and is given by

$$\mu_r = E\left[(X - \mu)^r\right], \qquad r = 0, 1, 2, \ldots$$
$$= \sum (X - \mu)^r P(X = x) \text{ for discrete random variable.}$$

$$\int_{-\infty}^{\infty} (x-\mu)^r f(x)\,dx \quad \text{for continuous random variable.}$$

where $f(x)$ is the pdf of X.
When $r = 0$, $\mu_0 = 1$,

$$r = 1,\ \mu_1 = E[(X-\mu)] = E(X) - \mu = 0$$
$$\therefore \mu_1 = 0$$
$$\mu_2 = E[(X-\mu)^2]$$

The second central moment is called variance.

Raw Moments

The r^{th} raw moment is the r^{th} moment of X about the origin, is defined as

$$\mu_r' = E(X^r), \quad r = 0, 1, 2, \ldots$$

We have,

$$\text{for } r = 0,\ \mu_0' = E(X^0) = E(1) = 1$$
$$\text{for } r = 1,\ \mu_1' = E(X) = \mu$$

Relationship Between Central Moments and Raw Moments

$$\mu_2 = E[(X-\mu)^2]$$
$$= E[X^2 + \mu^2 - 2\times\mu]$$
$$= E(X^2) + \mu^2 - 2E(X)\mu = E(X^2) + \mu^2 - 2\mu^2$$
$$\mu^2 = \mu_2' - \mu^2 \quad \text{Since } E(X) = \mu.$$
$$\mu_3 = E[(X-\mu)^3]$$
$$= E[X^3 - 3X^2\mu + 3\times\mu^2 - \mu^3]$$
$$= E(X^3) - 3E(X^2)\mu + 3\mu^2 E(X) - \mu^3$$
$$= \mu_3' - 3\mu_2'\mu + 3\mu^3 - \mu^3$$
$$\mu_3 = \mu_3' - 3\mu_2'\mu + 2\mu^3$$

Similarly $\mu_4 = \mu_4' - 4\mu_3'\mu + 6\mu_2'\mu^2 - 3\mu^4$
In general,

$$\mu_r = \mu_r' - r_{C_1}\mu_{r-1}'\mu + \cdots (-1)^j r_{C_j}\mu_{r-j}'\mu^j + \cdots + \cdots (-1)^r \mu_0'\mu^r$$

The r^{th} moment of a variable X about any $x = A$ is also denoted by μ_r'

$$\therefore \mu_r' = E[(X-A)^r]$$

EXAMPLE 3.14

Find the first four moments

(i) About the origin

(ii) About the mean for a random variable X with density function

$$f(x) = \begin{cases} \dfrac{4x(9-x^2)}{81}, & 0 \le x \le 3 \\ 0, & \text{otherwise} \end{cases}$$

Solution:

$$\mu_1' = E(X) = \frac{4}{81}\int_0^3 x \cdot x(9-x^2)\,dx$$

$$= \frac{4}{81}\int_0^3 (9x^2 - x^4)\,dx = \frac{4}{81}\left[9\frac{x^3}{3}\Big|_0^3 - \frac{x^5}{4}\Big|_0^3\right]$$

$$= \frac{4}{81}\left[3(27) - \frac{(81)^3}{5}\right] = \frac{4(81)}{81}\left[1 - \frac{3}{5}\right]$$

$$\mu_1' = \frac{8}{5} = \mu$$

$$\mu_2' = E(X^2) = \frac{4}{81}\int_0^3 x^2 \cdot x(9-x^2)\,dx$$

$$= \frac{4}{81}\int_0^3 (9x^3 - x^5)\,dx = \frac{4}{81}\left[9\frac{x^4}{4} - \frac{x^6}{6}\Big|_0^3\right]$$

$$= \frac{4}{81}\left[\frac{9}{4}(81) - \frac{81\times 9}{6}\right] = \frac{4(9\times 81)}{81}\left[\frac{1}{4} - \frac{1}{6}\right]$$

$$= 4\times 9\left(\frac{3-2}{12}\right) = 3$$

$$\mu_3' = E(X^3) = \frac{4}{81}\int_0^3 x^3 \cdot x(9-x^2)\,dx$$

$$= \frac{4}{81}\int_0^3 (9x^4 - x^6)\,dx = \frac{4}{81}\left[\frac{9x^5}{5} - \frac{x^7}{7}\Big|_0^3\right]$$

$$= \frac{4}{18}\left[\frac{9}{5}(81\times 3) - \frac{1}{7}(81\times 27)\right] = \frac{4}{35}\left[\frac{9(7)\times 3 - 5\times 27}{1}\right]$$

$$= \frac{4\times 27\times 2}{35} = \frac{216}{35}$$

$$\mu_4' = E(X^4) = \frac{4}{81}\int_0^3 x^4 \cdot x(9 - x^2)\, dx$$

$$= \frac{4}{81}\int_0^3 (9x^5 - x^7)\, dx = \frac{4}{81}\left[\left|9\frac{x^6}{6} - \frac{x^8}{8}\right|_0^3\right]$$

$$= \frac{4 \times 81}{81}\left[\frac{9 \times 9}{6} - \frac{81}{8}\right] = \frac{4 \times 81^{27}}{24_{8}}(4 - 3)$$

$$= \frac{27}{2}$$

Now obtaining the central moments using the result,

$$\mu_2 = \mu_2' - \mu^2$$

$$\mu_2 = 3 - \left(\frac{8}{5}\right)^2 = 3 - \frac{64}{25} = \frac{11}{25} = \sigma^2$$

$$\mu_3 = \mu_3' - 2\mu_2'\mu + 2\mu^3$$

$$= \frac{216}{35} - 3(3)\left(\frac{8}{5}\right) + 2\left(\frac{8}{5}\right)^3$$

$$= \frac{-32}{875}$$

$$\mu_4 = \mu_4' - 4\mu_3'\mu + 6\mu_2'\mu^2 - 3\mu^4$$

$$= \frac{27}{2} - 4\left(\frac{216}{35}\right)\left(\frac{8}{5}\right) + 6(3)\left(\frac{8}{5}\right)^2 - 3\left(\frac{8}{5}\right)^4$$

$$\therefore \mu_4 = \frac{3693}{8750}$$

3.9 MOMENT GENERATING FUNCTION

The main purpose of moment generating function (MGF) is to determine (generate) the moments of distribution.

The MGF of a random variable X (about origin) is given by

$$M_X(t) = E[e^{tx}] = \sum e^{tx} f(x) \text{ for a discrete } X.$$

$$= \int_{-\infty}^{\infty} e^{tx} f(x)\, dx \quad \text{for continuous } X.$$

$$E[e^{tx}] = E\left[1 + tx + \frac{t^2X^2}{2!} + \frac{t^3X^3}{3!} + \ldots + \frac{t^rX^r}{r!} + \ldots\right]$$

$$= 1 + tE(X) + \frac{t^2}{2!}E(X^2) + \frac{t^3}{3!}E(X^3) + \ldots \frac{t^r}{r!}E(X^r) + \ldots$$

$$E[e^{tx}] = 1 + t\mu_1' + \frac{t^2}{2!}\mu_2' + \ldots + \frac{t^r}{r!}\mu_r' + \ldots$$

$$\mu_X(t) = \sum_{r=0}^{\infty} \frac{t^r}{r!}\mu_r'$$

Differentiating the above equation w.r.t t for r times and putting $t = 0$ we get,

$$\mu_r' = \left|\frac{d^r}{dt^r}(M_X(t))\right|_{t=0}$$

Worked Out Example

EXAMPLE 3.15

A random variable X has the discrete uniform distribution $f(x, K) = \left\{\frac{1}{K}, \quad x = 1, 2, 3 \ldots K\right.$

Find the MGF of X.

Solution:

$$M_X(t) = \sum_{x=0}^{\infty} e^{tx} f(x)$$

$$= 1 + \sum_{x=1}^{K} e^{tx} f(x) + \sum_{x=K+1}^{\infty} e^{tx} f(x)$$

$$= \sum_{x=1}^{K} e^{tx} . \frac{1}{K} + \sum e^{tx} 0$$

$$= \sum_{x=1}^{K} e^{tx} f(x)$$

$$= \frac{1}{K}\left[e^t + e^{2t} + \ldots + e^{Kt}\right]$$

$$= \frac{1}{K}\left[e^t(1 + e^t + \ldots + e^{(K-1)t})\right]$$

$$= \frac{e^t}{K}\frac{(1 - e^{Kt})}{1 - e^t}$$ Since the above series is a G.P., $S_n = \frac{a(1 - r^n)}{1 - r}$

$$\therefore M_X(t) = \frac{e^t(1 - e^{Kt})}{K(1 - e^t)}$$

Caution:

- The MGF of X about the point $X = a$ is defined as,

$$M_X(t)(\text{about } X = a) = E[e^{t(X-a)}]$$
$$= E\left[1 + t(X-a) + \frac{t^2}{2!}(X-a)^2 + \ldots\right]$$
$$= 1 + t\mu_1' + \frac{t^2}{2!}\mu_2' + \ldots + \frac{t^r}{r!}\mu_r' + \ldots$$

where $E[e^{t(X-a)}] = \mu_r'$ is the r^{th} moment about any point $x = a$.

Properties of Moment Generating Functions

1. $M_{X+a}(t) = e^{at} M_X(t)$

   Proof:

$$M_X(t) = E[e^{tX}]$$
$$M_{X+a}(t) = E[e^{t(X+a)}]$$
$$= E[e^{tX} e^{at}]$$
$$\therefore M_{X+a}(t) = e^{at}\, E[e^{tX}] = e^{at} M_X(t)$$

2. $M_{aX}(t) = M_X(at)$

   Proof:

$$M_X(t) = E[e^{tX}]$$
$$M_{aX}(t) = E[e^{t(aX)}] = E[e^{a(tX)}]$$
$$= E[e^{at(X)}]$$
$$\therefore M_{aX}(t) = M_X(at)$$

3. If $X_1, X_2, \ldots X_n$ are independent random variables with MGFs $M_{X_1}(t), M_{X_2}(t) \ldots, M_{X_n}(t)$, respectively, then

$$M_{X_1+X_2+\ldots+X_n}(t) = M_{X_1}(t) \ldots M_{X_2}(t) \ldots M_{X_n}(t).$$

   Proof:

$$M_X(t) = E[e^{tX}]$$
$$M_{X_1+X_2+\ldots+X_n}(t) = [Ee^{t(X_1+X_2+\ldots+X_n)}]$$
$$= E[e^{tX_1+tX_2+\ldots+tX_n}]$$
$$= E[e^{tX_1+tX_2+\ldots tX_n}]$$
$$= E[e^{tX_1}]E[e^{tX_2}]\ldots E[e^{tX_n}]$$

   Since $X_1, X_2, \ldots X_n$ are independent random variables

$$E[XY] = E[X]E[Y]$$

$$\therefore M_{X_1+X_2+\ldots+X_n}(t) = M_{X_1}(t) \,.\, M_{X_2}(t) \ldots M_{X_n}(t)$$

4. Effect of change of origin and scale on MGF.

   Proof: Let X be a variable which transforms a new variable U.

   Let $U = \dfrac{X-a}{h}$

$$M_u(t) = E[e^{ut}] = E\left[e^{\left(\frac{X-a}{h}\right)t}\right] = E\left[e^{\frac{Xt}{h}} \cdot e^{-\frac{at}{h}}\right]$$
$$= e^{-\frac{at}{h}} E\left[e^{X\left(\frac{t}{h}\right)}\right] = e^{-\frac{at}{h}} M_X\left(\frac{t}{h}\right)$$

   Where $M_X(t)$ is the MGF of X about the origin.

5. **Uniqueness Theorem:**

   Let X and Y be two random variables with moment-generating functions $M_X(t)$ and $M_Y(t)$ respectively. If $M_X(t) = M_Y(t)$ for all values of t, then X and Y have the same probability distribution.

3.10 CHARACTERISTIC FUNCTION

The characteristic function of a random variable X is given by,

$$\phi_X(t) = E[e^{itX}] = \sum_x e^{itx} f(x) \qquad \text{if } X \text{ is discrete}$$

$$= \int_{-\infty}^{\infty} e^{itx} f(x)\, dx \qquad \text{if } X \text{ is continuous}$$

$$\phi_X(t) = E[e^{itX}] = E\left[1 + itX + \frac{(it)^2}{2!}X^2 + \ldots + \frac{(it)^r}{r!}X^r + \ldots\right]$$

$$= 1 + itE(X) + \frac{(it)^2}{2!}E(X^2) + \ldots + \frac{(it)^r}{r!}E(X^r) + \ldots$$

$$= 1 + it\mu_1' + \frac{(it)^2}{2!}\mu_2' + \ldots + \frac{(it)^r}{r!}\mu_r^1 + \ldots$$

where $\mu_r' = E[X^r]$ is r^{th} moment of X about origin.

$$\phi_X(t) = \Sigma \frac{(it)^r}{r!}\mu_r'$$

$$\therefore \mu_r' = \text{Coefficient of } \frac{(it)^r}{r!} \text{ in } \phi_X(t)$$

Properties of Characteristic Function

1. $\phi_{CX}(t) = \phi_X(Ct)$, C being constant.

   Proof: $\phi_X(t) = E[e^{itX}]$

$$\phi_{CX}(t) = E[e^{it(CX)}]$$
$$= E[e^{i(Ct)X}] = \phi_X(Ct)$$

2. If X_1 and X_2 are independent random variables, then

$$\phi_{X_1+X_2}(t) = \phi_{X_1}(t) \cdot \phi_{X_2}(t)$$

   Proof: $\phi_X(t) = E[e^{itX}]$

$$\phi_{X_1+X_2}(t) = E[e^{it(X_1+X_2)}]$$
$$= E[e^{itX_1} \cdot e^{itX_2}]$$
$$= \phi_{X_1}(t) \cdot \phi_{X_2}(t)$$ Since X and Y are independent random variables.

3. Effect of change of origin and scale on characteristic function.

   Proof: If $U = \dfrac{X-a}{h}$ is the new variable and a and h are constants then $\phi_U(t) = E[e^{iUt}]$

$$= E\left[e^{it\left(\frac{X-a}{h}\right)}\right] = E\left[e^{\frac{iXt}{h}} \cdot e^{\frac{-i^{at}}{h}}\right]$$
$$= e^{\frac{-i^{at}}{h}} E\left[e^{i\left(\frac{t}{h}\right)X}\right] = e^{\frac{-i^{at}}{h}} \phi_X\left(\frac{t}{h}\right)$$

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

This represents the Fourier transform of the density function $f(x)$. We can easily determine the density function from the characteristic function.

(i.e.,) $f(x) = \dfrac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itx} \phi_X(t)\, dt$

which is the inverse Fourier transform.

This is the reason why characteristic functions have been introduced. In addition, another reason for the usage of characteristic function is that it always exists whereas the MGF may not exist.

Worked Out Examples

EXAMPLE 3.16

Find the characteristic function of the random variable X having the density function given by

$$f(x) = \begin{cases} \dfrac{1}{2a}, & |x| < a \\ 0, & \text{otherwise} \end{cases}$$

Solution: The characteristic function of X is given by

$$\phi_X(t) = \int_{-\infty}^{\infty} \frac{1}{2a} e^{itX} f(x)\, dx$$

$$= \int_{-a}^{a} e^{itx} \cdot \frac{1}{2a}\, dx = \frac{1}{2a} \frac{e^{itx}}{it}\Big|_{-a}^{a}$$

$$= \frac{1}{it\,2a}\left[e^{iat} - e^{-iat}\right]$$

$$= \frac{1}{at}\left[\frac{e^{iat} - e^{-iat}}{2i}\right]$$

$$\phi_X(t) = \frac{\sin at}{at}$$

EXAMPLE 3.17

Find the characteristic function of the random variable with density function,

$$f(x) = \begin{cases} \frac{x}{2}, & 0 \le x \le 2 \\ 0, & \text{otherwise} \end{cases}$$

Solution: The characteristic function of the random variable X is given by

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

$$= \int_{0}^{2} e^{itx} \cdot \frac{x}{2}\, dx$$

$$= \frac{1}{2}\int_{0}^{2} e^{itx} \cdot x\, dx$$

$$= \frac{1}{2}\left[x \frac{e^{itx}}{it}\Big|_0^2 - \frac{1}{it}\int_0^2 e^{itx}\, dx\right]$$

$$= \frac{1}{2}\left[\frac{2e^{2it}}{it} - \frac{1}{it}\left|\frac{e^{itx}}{it}\right|_0^2\right]$$

$$= \frac{1}{2}\left[\frac{2e^{2it}}{it} - \frac{1}{(it)^2}(e^{2it} - 1)\right]$$

$$= \frac{1}{2(it)^2}\left[2ite^{2it} - e^{2it} - 1\right]$$

Work Book Exercises

6. Let X be a random variable with probability distribution

$$f(x)=\begin{cases}\frac{1}{3}, & x=1,2,3\\ 0, & \text{elsewhere}\end{cases}$$

Find the probability distribution of $Y=2X-1$.

7. Let X be a random variable with the following probability distribution:

X	–2	3	5
$f(x)$	0.3	0.2	0.5

(i) Find $E(X)$, $V(X)$.

(ii) Find the means and variance of the random variable $Z=5X+3$.

8. Two random variables X and Y take the values 1, 2, 3 along with the probabilities shown below:

Y \ X	1	3	9
1	K	K	$2K$
2	K	$2K$	K
3	$2K$	K	$2K$

(i) Find K.

(ii) Find $E(X+Y)$.

(iii) Find Var(X), Cov(X,Y).

(iv) Find the conditional distribution of Y given that $X=3$.

9. Find the following:

(i) Var(X), Var(Y)

(ii) The conditional distribution of X given $Y=4$.

(iii) The conditional distribution of Y given $X=-2$.

X \ Y	1	4
–2	0	$\frac{1}{4}$
–1	$\frac{1}{4}$	0
1	$\frac{1}{4}$	0
2	0	$\frac{1}{4}$

10. Suppose a random variable X has mean $\mu = 25$ and standard deviation $\sigma = 2$. Use Chebychev's inequality to find $P(X \le 35)$ as $P(X \ge 20)$.

[**Ans.:** 0.96, 0.84]

11. Consider the joint distribution of X and Y whose probabilities are as follows:

X \ Y	–4	2	7
1	$\frac{1}{8}$	$\frac{1}{4}$	$\frac{1}{8}$
5	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{8}$

(i) Find $E(X)$, $E(Y)$.

(ii) Find Var(X), Var(Y).

(iii) Find conditional distribution of Y given X and Y.

12. Suppose X and Y have the following joint probability distribution:

$f(x, y)$	x	2	4
	1	0.10	0.15
y	3	0.20	0.30
	5	0.10	0.15

(i) Find marginal distributions of X and Y.

(ii) Find $E(X)$, $V(X)$.

(iii) Find $E(Y)$, Var(Y).

(iv) Find $P\left(\frac{Y=3}{X=2}\right)$.

13. The joint density function of X and Y is given as

$$f(x,y) = \begin{cases} \frac{6-x-y}{8}, & 0 \le x < 2, \quad 0 < y < 4 \\ 0, & \text{elsewhere} \end{cases}$$

(i) Find the marginal distributions of X and Y.

(ii) Find $E(X+Y)$.

(iii) Find $P\left(\frac{1<Y<3}{X=2}\right)$.

14. Find the expected value of $z = \sqrt{X^2 + Y^2}$ given the joint density of X and Y as

$$f(x,y) = \begin{cases} 4xy, & 0 < x < 1, \quad 0 < y < 1 \\ 0, & \text{elsewhere} \end{cases}$$

15. Find the MGF of the binomial random variable X.

16. Let X_1 and X_2 be independent random variables each having probability distribution.

$$f(x) = \begin{cases} e^{-x}, & x > 0 \\ 0, & \text{elsewhere} \end{cases}$$

Show that the random variables Y_1 and Y_2 are independent when $Y_1 = X_1 + X_2$ and $Y_2 = \dfrac{X_1}{(X_1 + X_2)}$

17. (i) Find MGF of the random variable

X	$\frac{1}{2}$	$-\frac{1}{2}$
$P(X = x)$	$\frac{1}{2}$	$\frac{1}{2}$

(ii) In addition find first four moments about the origin.

18. Find the MGF and characteristic function of a random variable with the following density function:

$$f(x) = \begin{cases} \dfrac{x}{2}, & 0 \le x \le 2 \\ 0, & \text{otherwise} \end{cases}$$

19. Two random variables X and Y have the following joint probability density function:

$$f(x, y) = \begin{cases} K(u - x - y), \\ 0 \le x \le 2, \\ 0 \le y \le 2 \\ 0, \text{otherwise} \end{cases}$$

Find the following:

(i) Constant K

(ii) Marginal density function of X and Y

(iii) Conditional density functions

(iv) Var(X), Var(Y), and Cov(x, y)

20. Find the following:

(i) $E\left(\dfrac{Y}{X} = x\right)$

(ii) Cov(x, y) when $f(x, y) = 3(x + y)$, $0 \le x \le 1$, $0 \le y \le 1$, $0 \le x$ 2y ≤ 1

DEFINITIONS AT A GLANCE

Mathematical Expectation: Let X be a random variable with probability mass function $P(X = x) = f(x)$. Then the mathematical expectation of X is given by

$$E(X) = \sum_x xP(X = x); \quad \text{if } X \text{ is discrete}$$

$$= \int_{-\infty}^{\infty} xf(x)\,dx; \quad \text{if } X \text{ is continuous.}$$

If X, Y are two random variables then

$$E(X+Y) = E(X) + E(Y)$$
$$E(XY) = E(X)E(Y), \text{ provided } X, Y \text{ are independent.}$$
$$E(aX+b) = aE(X) + b$$

Variance: If X is a random variable then variance of X is given by

$$V(X) = E(X^2) - [E(X)]^2$$

If X is a random variable, then

$$V(aX+b) = a^2V(X)$$
$$V(aX) = a^2V(X)$$
$$V(X+b) = V(X)$$

Chebychev's theorem: If X is a random variable with mean μ and variance σ^2 then for any positive number K we have

$$P\{|X-\mu| < K\sigma\} > 1 - \frac{1}{K^2}$$

or $$P\{|X-\mu| \geq K\sigma\} \leq \frac{1}{K^2}$$

FORMULAE AT A GLANCE

- Expectation of a function of a random variable, $E(g(X)) = \int_{-\infty}^{\infty} g(x)\, f(x)\, dx$ where $f(x)$ is the marginal density of X.

  $E(g(X,Y)) = \int_{-\infty}^{\infty} g(x,y)\, f(x,y)\, dxdy$ where $f(x, y)$ is the joint density function of the random variables X and Y.
- Variance for joint distributions is given by
  $V(X) = E(X^2) - [E(X)]^2$ where

$$E(X) = \int_{-\infty}^{\infty} xg(x)\, dx$$

$$E(X^2) = \int_{-\infty}^{\infty} x^2 g(x)\, dx \text{ and}$$

$$g(X) = \int_{-\infty}^{\infty} f(x,y)\, dy \text{ is the marginal density of the random variable } X.$$

- Covariance between the random variables X and Y is given by

$$\text{Cov}(X, Y) = E\{[X - E(X)][Y - E(Y)]\}$$
$$= E(XY) - E(X)E(Y)$$

- $\text{Cov}(aX, bY) = ab\text{Cov}(X, Y)$
  $\text{Cov}(X + a, Y + b) = \text{Cov}(X, Y)$
  $\text{Cov}(aX + b, cY + d) = ac\text{Cov}(X, Y)$
- The conditional expectation or conditional mean of Y on X is given by $E\left(\frac{Y}{X}\right) = \int_{-\infty}^{\infty} y f\left(\frac{Y}{X}\right) dy$ where the conditional distribution function is given by $f\left(\frac{y}{x}\right) = \frac{f(x,y)}{g(x)}$ and $f(x, y)$ is the joint distribution function and $g(x)$ is the marginal density function of X.
- The r^{th} central moments about mean is given by

$$\mu_r = E[(X - \mu)^r]$$
$$= \int_{-\infty}^{\infty} (x - \mu)^r f(x)\ dx$$

- The r^{th} raw moment which is about the origin is given by

$$\mu_r' = E(X^r)$$
$$= \int_{-\infty}^{\infty} x^r f(x)\ dx$$
$$\mu_2 = \mu_2' - \mu_1'^2$$
$$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2\mu_1'^3$$
$$\mu_4 = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'\mu_1'^2 - 3\mu_1'^4$$

- If X is a continuous random variable, then the moment generating function of X is given by

$$M_X(t) = \text{MGF of } X$$
$$= E[e^{tx}] = \int_{-\infty}^{\infty} e^{tx} f(x)\, dx.$$

- The relation between the raw moments and the MGF of X is given by

$$M_X(t) = \sum_{r=0}^{\infty} \frac{t^r}{r!} \mu_r'.$$

- If X is a random variable and 'a' is constant then $M_{X+a}(t) = e^{at} M_X(t)$

$$M_{aX}(t) = M_X^{(at)}$$
$$M_{X_1 X_2}(t) = M_{X_1}(t) \cdot M_{X_2}(t) \text{ if } X_1, X_2 \text{ are independent random variables.}$$

- The characteristic function of a random variable X is given by $\phi_X(t) = \int_{-\infty}^{\infty} e^{itx} f(x)\, dx.$

- The relation between $\phi_X(t)$ and the raw moments (moments about the origin) is given by

$$\phi_X(t) = \sum \frac{(it)^r}{r!}\mu_r' .$$

- If C is a Constant ad $\phi_X(t)$ is the characteristic function of X then

$$\phi_{CX}(t) = \phi_X(Ct)$$

$\phi_{X_1+X_2}(t) = \phi_{X_1}(t) \cdot \phi_{X_2}(t)$ if X_1, X_2 are independent random variables.

OBJECTIVE TYPE QUESTIONS

1. Two random variables X and Y are such that $E(XY) = E(X)E(Y)$ only if
   (a) X and Y are identical
   (b) X and Y are stochastically independent
   (c) for all X and Y
   (d) none

2. $V(X + C) =$ _____ where C is a constant.
   (a) $V(X)$
   (b) $CV(X)$
   (c) $C^2V(X)$
   (d) none

3. If X and Y are independent $V(aX \pm by) =$ _____ where a and b are constants.
   (a) $a^2V(X) \pm b^2V(Y)$
   (b) $aV(X) \pm bV(Y)$
   (c) $V(X) + V(Y)$
   (d) none

4. $\text{Cov}(X + a, Y + b) =$ _____.
   (a) $a\text{Cov}(X, Y)$
   (b) $b\text{Cov}(X, Y)$
   (c) $\text{Cov}(X, Y)$
   (d) none

5. $\text{Cov}(aX + b, cY + d) =$ _____.
   (a) $ac\text{Cov}(X, Y)$
   (b) $bd\text{Cov}(X, Y)$
   (c) $\text{Cov}(X, Y)$
   (d) none

6. $E[\psi(X) + a] =$ _____.
   (a) $E[\psi(X)]$
   (b) $E[\psi(X)] + a$
   (c) $E[\psi(X)] + a$
   (d) none

7. $E[g(X)] > g[E(X)]$ only if
   (a) g is a continuous function
   (b) g is a convex function
   (c) both (a) and (b)
   (d) none

8. A continuous function $g(x)$ has interval such that $g\left(\frac{x_1 + x_2}{2}\right) \le \frac{1}{2}g(x_1) + \frac{1}{2}g(x_2)$ then
   (a) g is convex
   (b) continuous but not convex
   (c) g is odd
   (d) none

9. $|E(XY)|^2 \le E(X^2)E(Y^2)$. This relation is
   (a) Jensen's inequality
   (b) Cauchy Schwartz inequality
   (c) Concave function inequality
   (d) none

10. If $V(X) = 1$, then $V(2X \pm 3)$ is

(a) 5
(b) 13
(c) 4
(d) none

11. X and Y are independent then

(a) $\text{Cov}(X, Y) = 1$
(b) $\text{Cov}(X, Y) = 0$
(c) $E(XY) = E(X)E(Y)$
(d) both (b) and (C)
(e) none

12. $E(X - K)^2$ is minimum if

(a) $K < E(X)$
(b) $K > E(X)$
(c) $K = E(X)$
(d) both (a) and (c)

ANSWERS

1. (b) 2. (a) 3. (a) 4. (c) 5. (a) 6. (c) 7. (c) 8. (a)
9. (b) 10. (c) 11. (b) 12. (c)

4 Standard Discrete Distributions

Prerequisites

Before you start reading this unit, you should:

- Be thorough in finding probability mass functions and its properties
- Be able to find certain probabilities

Learning Objectives

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

- Understand the concepts of discrete distributions
- Differentiate between the various distributions and use suitable distribution in numerical problems
- Find moments of any discrete distributions
- Find moment generating functions for the above discrete distributions

INTRODUCTION

In this unit, we deal with special and standard distributions which are discrete distributions. Some of the discrete probability distributions that are successfully applied in a variety of decision making situations are introduced. Here, the parameters, that is, the quantities that are constants for particular distributions, but can take on different values for different members of distributions of the same kind. Now, let us start studying the well-known Binomial distribution with its properties and the name given to it and some other distributions followed by it.

4.1 BINOMIAL DISTRIBUTION

This is a discrete distribution. Suppose an experiment is conducted under essentially identical conditions n times where these are independent trials. Let 'p' be the probability of occurrence of an event called the probability of success and 'q' its complementary event is called the probability of failure. Then the probability of getting exactly r successes in n independent trials is $nC_r p^r q^{n-r}$.

Definition

A random variable X has a Binomial distribution if X assumes only non-negative values and its probability distribution is given by

$$P(X = r) = \begin{cases} nC_r p^r q^{n-r}, & r = 0,1,...n \\ 0, & \text{otherwise} \end{cases} \tag{4.1}$$

Caution:

- n and p are called the parameters and hence we use the notation $X \sim B(n, p)$ to say that X follows Binomial distribution with n and p as parameters.

The assumptions followed by a binomial variate are as follows:

- There are only two possible outcomes for each trial called success and failure.
- The probability of success in each trial is constant.
- The number of trials are independent.
- The probability distribution given by equation (4.1) is a probability mass function.

$$\begin{aligned}\sum P(X=x) &= \sum_{x=0}^{n} nC_x p^x q^{n-x} \\ &= nC_0 p^0 q^{n-0} + nC_1 p^1 q^{n-1} + \ldots nC_n p^n q^{n-n} \\ &= (p+q)^n \\ &= 1 \\ \sum P(X=x) &= 1\end{aligned}$$

- $\sum nC_x p^x q^{n-x} = \sum n-1C_{x-1} p^{x-1} q^{n-1-(x-1)} = \sum n-2C_{x-2} p^{x-2} q^{n-2-(x-2)} = 1$

Moments of Binomial Distribution

Mean of Binomial Distribution

$$\begin{aligned}E(X) &= \sum_{x=1}^{n} x\, P(X=x) \\ &= \sum_{x=1}^{n} x.nC_x p^x q^{n-x} \\ &= \sum x \frac{n!}{x!(n-x)!} p\, p^{x-1} q^{n-x} \\ &= np \sum \frac{(n-1)!}{(x-1)!\,[n-1-(x-1)]!} p^{x-1} q^{n-1(x-1)} \\ &= np \sum n-1C_{x-1} p^{x-1} q^{n-1(x-1)} \\ E(X) &= np\end{aligned}$$

$$\begin{aligned}E(X^2) &= \sum x^2 nC_x p^x q^{n-x} \\ &= \sum [x(x-1)+\mathrm{x}] nC_x p^x q^{n-x} \\ &= \sum x(x-1) nC_x p^x q^{n-x} + \sum x nC_x p^x q^{n-x} \\ &= \sum x(x-1) \frac{n!}{x!(n-x)!} p^x q^{n-x} + np \\ &= \sum x(x-1) \frac{n(n-1)(n-2)!}{x(x-1)(x-2)![n-2-(x-2)]!} p^2 p^{x-2} q^{n-2-(x-2)} + np \\ &= n(n-1)p^2 \sum \frac{(n-2)!}{(x-2)![n-2-(x-2)]!} p^{x-2} q^{n-2-(x-2)} + np \\ E(X^2) &= n(n-1)p^2 + np\end{aligned}$$

Variance of Binomial Distribution

The variance of a distribution is given by

$$\begin{aligned} V(X) &= E(X^2) - [E(X)]^2 \\ &= n(n-1)p^2 + np - n^2p^2 \\ &= n^2p^2 - np^2 + np - n^2p^2 \\ &= np(1-p) \\ &= npq \end{aligned}$$

$$\boxed{V(X) = npq}$$

Caution:

- The mean of Binomial distribution is *np* and variance of Binomial distribution is *npq.*
- The number of trials '*n*' is finite.
- The problems relating to tossing of a coin or throwing a die or drawing cards from a pack of cards with replacement lead to Binomial distribution.

Worked Out Examples

EXAMPLE 4.1

In a certain district, the need for money to buy drugs is stated as the reason for 75% of all thefts. Find the probability that among the next 5 theft cases reported in this district,

(i) Exactly 2 resulted from the need for money to buy drugs.

(ii) At most 3 resulted from the need for money to buy drugs.

Solution: Let X denote the need for money to buy drugs in a certain district. Let p denote the percentage of thefts in the city.

$$\begin{aligned} p &= 0.75, \\ q &= 1 - p \\ &= 1 - 0.75 \\ &= 0.25 \end{aligned}$$

This represents a Binomial distribution.

$$\begin{aligned} P(X = x) &= \text{Probability that there are exactly } x \text{ successes} \\ &= nC_x p^x q^{n-x} \end{aligned}$$

(i) $P(X = 2)$ = Probability that exactly 2 resulted from the need for money to buy drugs

$$\begin{aligned} &= 5C_2 p^2 q^{5-2} \\ &= 10(0.75)^2 (0.25)^3 \\ &= 0.0878 \end{aligned}$$

(ii) $P(X \le 3)$ = Probability that at most 3 resulted from the need for money to buy drugs

$$P(X \le 3) = P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3)$$
$$= 5C_0 p^0 q^{5-0} + 5C_1 p^1 q^{5-1} + 5C_2 p^2 q^{5-2} + 5C_3 p^3 q^{5-3}$$
$$= (0.25)^5 + 5\,(0.75)\,(0.25)^4 + 10(0.75)^2(0.25)^3 + 10(0.75)3(0.25)^2$$
$$= 0.0009765 + 0.014648 + 0.08789 + 0.263671$$
$$= 0.367186$$

EXAMPLE 4.2

Six cards are drawn from a pack of 52 cards. Find the probability that there are

(i) At least three diamonds

(ii) None is a diamond (JNTU Aug./Sep. 2008 set-1)

Solution: Let X denote a diamond card from the 52 pack of cards.

$P(X = x)$ = Probability that there are exactly x successes

$$= nC_x p^x q^{n-x}$$

(i) p = Probability of drawing a diamond $= \frac{1}{13}$

$q = 1 - p$

$$= 1 - \frac{1}{13}$$

$q = \frac{12}{13}$, $n = 6$

$P(X \ge 3)$ = Probability that there are at least three diamonds

$$P(X \ge 3) = 1 - P(X < 3)$$
$$= P(X = 0) + P(X = 1) + P(X = 2)$$
$$= 6C_0 p^0 q^{6-0} + 6C_1 p^1 q^{6-1} + 6C_2 p^2 q^{6-2}$$
$$= \left(\frac{12}{13}\right)^6 + 6\left(\frac{1}{13}\right)\left(\frac{12}{13}\right)^5 + 15\left(\frac{1}{13}\right)^2\left(\frac{12}{13}\right)^4$$
$$= 0.\,992377$$

(ii) Probability that none is a diamond $= P(X = 0)$

$$P(X = 0) = 6C_0 p^0 q^{6-0}$$
$$= \left(\frac{12}{13}\right)^6$$
$$= 0.618625$$

EXAMPLE 4.3

Two dice are thrown five times. If getting a double is a success, find the probability that getting the success

(i) At least once

(ii) Twice (JNTU Aug./Sep. 2008, set-3)

Solution: Let X denote the numbers on the two dice. Let p = getting a double when two dice are thrown.

No. of favourable cases for getting a double = {(1, 1), (2, 2), (3, 3), (4, 4), (5, 5), (6, 6)}
Total number of cases = 36

$$P = \frac{6}{36} = \frac{1}{6}$$

$$q = 1 - p$$

$$= 1 - \frac{1}{6}$$

$$= \frac{5}{6}, \; n = 5$$

$P(X = x)$ = Probability that there are exactly x successes

$$= nC_x p^x q^{n-x}$$

(i) $P(X > 1)$ = Probability of getting success at least once

$$P(X \geq 1) = 1 - P(X < 1)$$

$$= 1 - P(X = 0)$$

$$= 1 - 5C_0 p^0 q^{5-0}$$

$$= 1 - \left(\frac{5}{6}\right)^5$$

$$= 1 - 0.401878$$

$$= 0.598122$$

(ii) $P(X = 2)$ = Probability of getting success exactly two times

$$P(X = 2) = 5C_2 p^2 q^{5-2}$$

$$= 10\left(\frac{1}{6}\right)^2 \left(\frac{5}{6}\right)^3$$

$$= 0.160751$$

EXAMPLE 4.4

Out of 800 families with 5 children each, how many would you expect to have?

(i) Three boys

(ii) At least one boy (JNTU Nov. 2006, set-2)

Solution: Let X denote the number of boys in a family.

Probability of a boy = Probability of a girl

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

$P(X = x)$ = Probability that there are exactly x successes

$$= nC_x p^x q^{n-x}$$

(i) Probability of exactly 3 boys = $P(X = 3)$

$$P(X = 3) = 5C_3 p^3 q^{5-3}$$

$$= 10\left(\frac{1}{2}\right)^3\left(\frac{1}{2}\right)^2$$
$$= 10(0.03125)$$
$$= 0.3125$$

Hence, out of 800 families with 5 children each, the number of families with 3 boys each

= Probability of exactly 3 boys × 800
= 800 (0.3125)
= 250

(ii) Probability of at least one boy = P($X \geq 1$)

$$P(X \geq 1) = 1 - P(X < 1)$$
$$= 1 - P(X = 0)$$
$$P(X \geq 1) = 1 - 5C_0 p^0 q^{5-0}$$
$$= 1 - \left(\frac{1}{2}\right)^5$$
$$= 1 - 0.03125$$
$$= 0.96875$$

EXAMPLE 4.5

The mean of Binomial distribution is 3 and the variance is $\frac{9}{4}$. Find the following:

(i) Value of n

(ii) $P(X \geq 7)$

(iii) $P(1 \leq X < 6)$ (JNTU Aug. /Sep. 2008 set-4)

Solution: Given that the mean of Binomial distribution = np

$$np = 3 \tag{4.2}$$

In addition, given that the variance of Binomial distribution = npq

$$npq = \frac{9}{4}$$
$$\frac{npq}{np} = \frac{\frac{9}{4}}{3}$$
$$q = \frac{3}{4},\ p = 1 - q$$
$$= 1 - \frac{3}{4}$$
$$= \frac{1}{4}$$

(i) From equation (4.2), $n\left(\frac{1}{4}\right) = 3$

$$n = 3 \times 4 = 12$$

(ii) $P(X \geq 7) = P(X=7) + P(X=8) + P(X=9) + P(X=10) + P(X=11) + P(X=12)$

$= 12C_7 p^7 q^{12-7} + 12C_8 p^8 q^{12-8} + 12C_9 p^9 q^{12-9} + 12C_{10} p^{10} q^{12-10} + 12C_{11} p^{11} q^{12-11}$

$+ 12C_{12} p^{12} q^{12-12}$

$$P(X \geq 7) = 792\left(\frac{1}{4}\right)^7\left(\frac{3}{4}\right)^5 + 495\left(\frac{1}{4}\right)^8\left(\frac{3}{4}\right)^4 + 220\left(\frac{1}{4}\right)^9\left(\frac{3}{4}\right)^3 + 66\left(\frac{1}{4}\right)^{10}\left(\frac{3}{4}\right)^2 + \left(\frac{1}{4}\right)^{11}\left(\frac{3}{4}\right)^1 + \left(\frac{1}{4}\right)$$

$= 0.014253$

$P(1 \leq X < 6) = P(X=1) + P(X=2) + P(X=3) + P(X=4) + P(X=5)$

$= 12C_1 p^1 q^{12-1} + 12C_2 p^2 q^{12-2} + 12C_3 p^3 q^{12-3} + 12C_4 p^4 q^{12-4} + 12C_5 p^5 q^{12-5}$

$$= 12\left(\frac{1}{4}\right)^1\left(\frac{3}{4}\right)^{11} + 66\left(\frac{1}{4}\right)^2\left(\frac{3}{4}\right)^{10} + 220\left(\frac{1}{4}\right)^3\left(\frac{3}{4}\right)^9 + 495\left(\frac{1}{4}\right)^4\left(\frac{3}{4}\right)^8 + 792\left(\frac{1}{4}\right)^5\left(\frac{3}{4}\right)^8$$

$= 0.888111$

EXAMPLE 4.6

The probability of a man hitting a target is $\frac{1}{3}$.

(i) If he fires five times, what is the probability of his hitting the target at least twice?

(ii) How many times must he fire so that the probability of his hitting the target at least once is more than 90%? (JNTU Aug./Sep. 2006 set-1)

Solution: Let n = number of trials = 5

p = probability of hitting a target = $\frac{1}{3}$

q = probability of not hitting the target = $1 - \frac{1}{3} = \frac{2}{3}$

(i) Probability of the man hitting the target at least twice = $P(X \geq 2)$

$$P(X \geq 2) = 1 - P(X < 2)$$

$$= 1 - \{P(X=0) + P(X=1)\}$$

$$= 1 - \left\{5C_0\left(\frac{1}{3}\right)^0\left(\frac{2}{3}\right)^5 + 5C_1\left(\frac{1}{3}\right)^1\left(\frac{2}{3}\right)^4\right\}$$

$$= 1 - \left\{\left(\frac{2}{3}\right)^5 + \left(\frac{5}{3}\right)\left(\frac{2}{3}\right)^4\right\}$$

$$= 1 - \left(\frac{2}{3}\right)^4 \left\{\frac{2}{3} + \frac{5}{3}\right\}$$

$$= 1 - \left(\frac{2}{3}\right)^4 \left(\frac{7}{3}\right)$$

$$= 0.539095$$

(ii) Given that probability of hitting the target at least once > 90%

$$P \text{ (at least once)} > 90\%$$

$$P(X \geq 1) > 90\%$$

$$1 - (X < 1) > 0.9$$

$$1 - P(X = 0) > 0.9$$

$$1 - nC_0 \left(\frac{1}{3}\right)^0 \left(\frac{2}{3}\right)^n > 0.9$$

$$1 - \left(\frac{2}{3}\right)^n > 0.9$$

The above equation is satisfied for $n = 6$

$$\text{since } 1 - \left(\frac{2}{3}\right)^6 = 0.9122 > 0.9$$

EXAMPLE 4.7

In testing a certain kind of truck tyre over a rugged terrain, it is found that 25% of the trucks fail to complete the test run without a blowout and of the next 15 trucks tested, find the probability that:

(i) From 3 to 6 have blowouts

(ii) Fewer than 4 have blowouts

(iii) More than 5 have blowouts

Solution: Let n = Number of trucks tested = 15

p = Probability of trucks failing to complete the test run

= 25% = 0.25

$q = 1 - p = 0.75$

This is a Binomial distribution. The probability distribution is given by,

$$P(X = x) = nC_x p^x q^{n-x}$$

$P(X = x)$ = Probability that there are exactly x blowouts

(i) Probability that there are between 3 and 6 trucks with blowouts = $P(3 < x < 6)$

$$P(3 < x < 6) = P(X = 4) + P(X = 5)$$

$$= 15C_4 (0.25)^4 (0.75)^{15-4} + 15C_5 (0.25)^5 (0.75)^{15-5}$$

$$= 15C_4 (0.25)^4 (0.75)^{11} + 15C_5 (0.25)^5 (0.75)^{10}$$

$$= (1365)(0.003906)(0.042235) + (3003)(0.000977)(0.056314)$$
$$= 0.390345$$

(ii) Probability that there are fewer than 4 blowouts = $P(x < 4)$

$$P(x < 4) = P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3)$$
$$= 15C_0(0.25)^0(0.75)^{15-0} + 15C_1(0.25)^1(0.75)^{15-1} + 15C_2(0.25)^2(0.75)^{15-2} + 15C_3(0.25)^3(0.75)^{15-3}$$
$$= 15C_0(0.25)^0(0.75)^{15} + 15C_1(0.25)^1(0.75)^{14} + 15C_2(0.25)^2(0.75)^{13} + 15C_3(0.25)^3(0.75)^{12}$$
$$= 0.013363 + 0.066817 + 0.155907 + 0.225199$$
$$= 0.461286$$

(iii) Probability that there are more than 5 blowouts = P($x > 5$)

$$P(x < 5) = 1 - P(X \le 5)$$
$$= 1 - \{P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4) + P(X = 5)\}$$
$$P(x > 5) = 1 - \{15C_0(0.25)^0(0.75)^{15} + 15C_1(0.25)^1(0.75)^{14} + 15C_2(0.25)^2(0.75)^{13} + 15C_3(0.25)^3(0.75)^{12} + 15C_4(0.25)^4(0.75)^{11} + 15C_5(0.25)^5(0.75)^{10}$$
$$= 1 - \{0.013363 + 0.066817 + 0.155907 + 0.225199 + 0.168899 + 0.165146\}$$
$$= 1 - 0.795331$$
$$= 0.204669$$

EXAMPLE 4.8

The mean and variance of a Binomial distribution are 4 and $\frac{4}{3}$, respectively. Find $P(X \ge 1)$.

(JNTU April/May. 2004, set-1)

Solution: Mean of Binomial distribution = np

Given that $np = 4$ (4.3)

Variance of Binomial distribution = npq

Given that $npq = \frac{4}{3}$ (4.4)

On dividing (4.4) by (4.3) we get,

$$\frac{npq}{np} = \frac{\frac{4}{3}}{4}$$
$$q = \frac{1}{3}$$
$$p = 1 - q = 1 - \frac{1}{3} = \frac{2}{3}$$
$$p = \frac{2}{3}$$

$$np = 4$$

$$n = \frac{4}{\frac{2}{3}} = 6$$

$$P(X \geq 1) = 1 - P(X < 1)$$

$$= 1 - P(X = 0)$$

$$P(X \geq 1) = 1 - 6C_0 \left(\frac{2}{3}\right)^0 \left(\frac{1}{3}\right)^6$$

$$= 1 - \left(\frac{1}{3}\right)^6$$

$$= 0.998628$$

EXAMPLE 4.9

Ten coins are thrown simultaneously. Find the probability of getting at least seven heads.

(JNTU April/May. 2004 set -2)

Solution: Let p = Probability of getting a head = $\frac{1}{2}$

q = Probability of getting a head = $\frac{1}{2}$

The probability of getting x heads in a throw of 10 coins = $P(X = x)$

$$P(X = x) = 10\,\mathrm{C}_x \left(\frac{1}{2}\right)^x \left(\frac{1}{2}\right)^{10-x}, \quad x = 0, 1, 2, \ldots 10$$

Probability of getting at least 7 heads when 10 coins are thrown = $P(X \geq 7)$

$$P(X \geq 7) = P(X = 7) + P(X = 8) + P(X = 9) + P(X = 10)$$

$$= 10C_7 \left(\frac{1}{2}\right)^7 \left(\frac{1}{2}\right)^3 + 10C_8 \left(\frac{1}{2}\right)^8 \left(\frac{1}{2}\right)^2 + 10C_9 \left(\frac{1}{2}\right)^9 \left(\frac{1}{2}\right)^1 + 10C_{10} \left(\frac{1}{2}\right)^{10} \left(\frac{1}{2}\right)^0$$

$$= 10C_7 \left(\frac{1}{2}\right)^{10} + 10C_8 \left(\frac{1}{2}\right)^{10} + 10C_9 \left(\frac{1}{2}\right)^{10} + 10C_{10} \left(\frac{1}{2}\right)^{10}$$

$$= \left(\frac{1}{2}\right)^{10} \{10C_7 + 10C_8 + 10C_9 + 10C_{10}\}$$

$$= \left(\frac{1}{2}\right)^{10} \{120 + 45 + 10 + 1\}$$

$$P(X \geq 7) = \frac{176}{1024} = 0.1719$$

EXAMPLE 4.10

The probability that a man aged 75 years will die before his next birthday is 0.031. What is the probability that out of 15 such men, at least 13 will reach their 76th birthday? (December'08, O.U)

Solution: Let p be the probability that a man aged 75 years will die before his next birthday.

$$p = 0.031,\ q = 1 - p = 1 - 0.031$$
$$= 0.969$$

Number of men, $n = 15$

$P(X \geq 13)$ = Probability that out of 15 such men, at least 13 will reach their 76th birthday

$$= P(X = 13) + P(X = 14) + P(X = 15)$$
$$= 15C_{13}(0.031)^{13}(0.969)^{2} + 15C_{14}(0.031)^{14}(0.969)^{1} + 15C_{15}(0.031)^{15}(0.969)^{0}$$
$$= 2.40735E - 18 + 1.10022E - 20 + 2.34653E - 23$$
$$= 2.41838E - 18 = 2.418 \times 10^{-18}$$

Work Book Exercises

1. There is 90% chance that a particular type of component performs adequately under high temperature conditions. Suppose the system has four such components, state the probability distributions that to compute the following probabilities and find the probability that
   (i) All the components perform adequately and hence the system is operative.
   (ii) The system is inoperative because more than two components fail. (April'99, O.U)
2. It is known that 60% of the mice inoculated with a serum are protected from a certain disease. If 5 mice are inoculated, find the probability that:
   (i) It contracts the disease
   (ii) Fewer than 2 contracts the disease
   (iii) More than 3 contracts the disease
3. Assume that 50% of all engineering students are good in mathematics. Determine the probability among 20 engineering students selected:
   (i) Exactly 8
   (ii) At least 9
   (iii) At most 7
   (iv) At least 2 and at most 10 are good in mathematics

Fitting of Binomial Distribution

A probability distribution is to be fit for a given set of data points. Here fitting of Binomial distribution is explained in detail. The main aim of fitting of a distribution is to find expected frequencies for a set of data points (x_i, f_i) which contain x_i and corresponding observed frequencies f_i. In order to find these we find the probabilities of the Binomial distribution using the recurrence relation given in the next section. Using these probabilities we find the expected frequencies,

$f(x) = NP(x)$, where $P(x)$ denotes the probabilities for $x = 0, 1, 2, \ldots n$. $N = \sum f_i$ where f_i denotes the observed frequencies of given data.

Recurrence Relation for the Probabilities of Binomial Distribution

The probability distribution for Binomial distribution is given by,

$$P(x) = P(X = x) = nC_x p^x q^{n-x}$$

$$P(x+1) = P(X = x+1) = nC_{x+1}\, p^{x+1} q^{n-x-1}$$

$$\frac{P(x+1)}{P(x)} = \frac{nC_{x+1} p^{x+1} q^{n-x-1}}{nC_x p^x q^{n-x}}$$

$$= \frac{\frac{n!}{(x+1)!(n-x-1)!} p^{x+1} q^{n-x-1}}{\frac{n!}{(x)!(n-x)!} p^x q^{n-x}}$$

$$p(x+1) = \frac{\frac{n!}{(x+1)x!(n-x-1)!} p^x \cdot p\, q^{n-x}}{\frac{n!}{x!(n-x)(n-x-1)!} p^x q^{n-x} \cdot q}$$

$$= \frac{n-x}{x+1} \cdot \frac{p}{q}$$

$$P(x+1) = \frac{n-x}{x+1} \cdot \frac{p}{q} P(x) \tag{4.5}$$

Moment Generating Function (MGF) of Binomial Distribution

Let X be a discrete random variable, which follows Binomial distribution. Then

$$M_x(t) = E(e^{tx}) = \sum_{x=0}^{n} e^{tx} nC_x p^x q^{n-x}$$

$$= \sum_{x=0}^{n} (pe^t)^x nC_x q^{n-x}$$

$$M_x(t) = (q + pe^t)^n$$

The MGF About Mean of Binomial distribution

Since the mean of Binomial distribution is np, we have

$$E\{e^{t(x-np)}\} = E\{e^{tx} e^{-tnp}\}$$

$$= e^{-tnp} E(e^{tx})$$

$$= e^{-tnp} M_x(t)$$

$$= e^{-tnp} (q + pe^t)^n$$

$$= (q + pe^t)^n (e^{-tp})^n$$

$$= (qe^{-tp} + pe^{t(1-p)})^n$$

$$= (qe^{-tp} + pe^{tq})^n$$

$$= \left[q\left\{1 - pt + \frac{p^2t^2}{2!} + \frac{p^3t^3}{3!} + \ldots\right\} + p\left\{1 + tq + \frac{q^2t^2}{2!} + \frac{q^2t^2}{3!}\right\}\right]^n$$

$$= \left[\{(q+p)\} + \left\{ pq\frac{t^2}{2!}(p+q) + pq\frac{t^3}{3!}(q^2 - p^2) + \ldots\right\}\right]^n$$

$$= \left[1 + nC_1\left\{ pq\frac{t^2}{2!}(p+q) + pq\frac{t^3}{3!}(q^2 - p^2) + \ldots\right\} + nC_2\left\{ pq\frac{t^2}{2!}(p+q)\right.\right.$$

$$\left.\left. + pq\frac{t^3}{3!}(q^2 - p^2) + ------\right\}^2 + ------\right] \quad \{\text{since } q + p = 1\}$$

From the above equation,

$$\mu_2 = \text{Coefficient of } \frac{t^2}{2!} = npq$$

$$\mu_3 = \text{Coefficient of } \frac{t^2}{3!} = npq(q - p)$$

Additive Property of Binomial Distribution

Let X and Y be two independent binomial variables with n_1, p_1 and n_2, p_2 as the parameters. Then the MGF's of X and Y are given by

$$M_x(t) = (q_1 + p_1e^t)^{n_1}$$

$$M_y(t) = (q_2 + p_2e^t)^{n_2}$$

Then $M_{x+y}(t) = M_x(t).\ M_y(t)$, since X and Y are independent.

$$= (q_1 + p_1e^t)^{n_1}.(q_2 + p_2e^t)^{n_2} \tag{4.6}$$

From Uniqueness theorem of MGF's it follows that $X + Y$ is not a binomial variate. Hence, the sum of two independent binomial variates is not a binomial variate. Therefore, Binomial distribution does not possess additive property.

Caution: Let $p_1 = p_2 = p$ then substituting in equation (4.6), we get

$$M_{x+y}(t) = (q + pe^t)^{n_1} \cdot (q + pe^t)^{n_2}$$

$$= (q + pe^t)^{n_1 + n_2}$$

This is the MGF of a binomial variate with parameters $(n_1 + n_2, p)$. Hence, the Binomial distribution possesses additive property only if $p_1 = p_2$.

Characteristic Function of Binomial Distribution

$$\varphi_x(t) = E(e^{itx}) = \sum_{x=0}^{n} e^{itx}\ nC_x p^x q^{n-x}$$

$$= \sum_{x=0}^{n} (pe^{it})^x\ nC_x q^{n-x}$$

$$= (q + pe^{it})^n$$

Worked Out Examples

EXAMPLE 4.11

A set of 6 similar coins are tossed 640 times with the following results:

Number of heads	0	1	2	3	4	5	6
Frequency	7	64	140	210	132	75	12

Fit a Binomial distribution to the above data assuming that the coins are unbiased.

Solution: The coin is unbiased:

Probability of a head, $p = \frac{1}{2}$

Probability of a tail, $q = \frac{1}{2}$

Number of coins, $n = 6$

The recurrence relation for probabilities of Binomial distribution is

$$P(x+1) = \left(\frac{n-x}{x+1}\right)\frac{p}{q}P(x)$$

$$\text{since } \frac{p}{q} = 1$$

$$P(x+1) = \left(\frac{n-x}{x+1}\right)P(x)$$

The expected frequencies are calculated using

$$f(x) = N \cdot P(x) \text{ where } N = 640$$

$$P(0) = 6C_0\left(\frac{1}{2}\right)^0\left(\frac{1}{2}\right)^6$$

$$= \left(\frac{1}{2}\right)^6 = \frac{1}{64}$$

$$f(0) = N \cdot P(0)$$

$$= 640\left(\frac{1}{64}\right) = 10$$

$$P(1) = \left(\frac{6-0}{0+1}\right)P(0)$$

$$P(1) = 6 \quad P(0) = \frac{6}{64}$$

$$f(1) = N \cdot P(1) = 640\left(\frac{6}{64}\right) = 60$$

$$P(2) = \left(\frac{6-1}{1+1}\right)P(1)$$

$$P(2) = \frac{5}{2}\ P(1) = \frac{5}{2}\frac{6}{64} = \frac{15}{64}$$

$$f(2) = N \cdot P(1) = 640\left(\frac{15}{64}\right) = 150$$

$$P(3) = \left(\frac{6-2}{2+1}\right)P(2)$$

$$P(3) = \frac{4}{3}P(2) = \left(\frac{4}{3}\right)\left(\frac{15}{64}\right) = \frac{5}{16}$$

$$f(3) = N \cdot P(3) = 640\left(\frac{5}{16}\right) = 200$$

$$P(4) = \left(\frac{6-3}{3+1}\right)P(3)$$

$$P(4) = \frac{3}{4}P(3) = \frac{3}{4}\left(\frac{5}{16}\right) = \frac{15}{64}$$

$$f(4) = N \cdot P(4) = 640\left(\frac{15}{64}\right) = 150$$

$$P(5) = \left(\frac{6-4}{4+1}\right)P(4)$$

$$P(5) = \frac{2}{5}P(4) = \left(\frac{2}{5}\right)\left(\frac{15}{64}\right) = \left(\frac{3}{32}\right)$$

$$f(5) = N \cdot P(5) = 640\left(\frac{3}{32}\right) = 60$$

$$P(6) = \left(\frac{6-5}{5+1}\right)P(5)$$

$$P(6) = \frac{1}{6}P(5) = \left(\frac{1}{6}\right)\left(\frac{3}{32}\right) = \left(\frac{1}{64}\right)$$

$$f(6) = N \cdot P(6) = 640\left(\frac{1}{64}\right) = 10$$

The expected frequencies of the data are as follows:

Number of heads	0	1	2	3	4	5	6	Total
Frequency	7	64	140	210	132	75	12	640
Expected frequency, $f(x)$	10	60	150	200	150	60	10	640

EXAMPLE 4.12

Seven coins are tossed and the number of heads are noted. The experiment is repeated 128 times and the following distribution is obtained:

Number of heads	0	1	2	3	4	5	6	7	Total
Frequency	7	6	19	35	30	23	7	1	128

Fit a Binomial distribution assuming that:

(i) The coin is unbiased

(ii) The nature of the coin is not known

Solution:

(i) The coin is unbiased:

Probability of a head, $p = \frac{1}{2}$

Probability of a tail, $q = \frac{1}{2}$

Number of coins, $n = 7$

The recurrence relation for probabilities of Binomial distribution is

$$P(x+1) = \left(\frac{n-x}{x+1}\right)\frac{p}{q}P(x)$$

since $\frac{p}{q} = 1$

$$P(x+1) = \left(\frac{n-x}{x+1}\right)P(x)$$

The expected frequencies are calculated using,
$f(x) = N \cdot P(x)$, where $N = 128$

$$P(0) = 7C_0\left(\frac{1}{2}\right)^0\left(\frac{1}{2}\right)^7$$

$$= \left(\frac{1}{2}\right)^7 = \frac{1}{128}$$

$$f(0) = N \cdot P(0)$$

$$= 128\left(\frac{1}{128}\right) = 1$$

$$P(1) = \left(\frac{7-0}{0+1}\right)P(0)$$

$$P(1) = 7P(0) = \frac{7}{128}$$

$$f(1) = N \cdot P(1) = 128\left(\frac{7}{128}\right) = 7$$

$$P(2) = \left(\frac{7-1}{1+1}\right)P(1)$$

$$P(2) = \frac{6}{2}P(1) = 3P(1) = \frac{21}{128}$$

$$f(2) = N \cdot P(2) = 128\left(\frac{21}{128}\right) = 21$$

$$P(3) = \left(\frac{7-2}{2+1}\right)P(2)$$

$$P(3) = \frac{5}{3}P(2) = \left(\frac{5}{3}\right)\left(\frac{21}{128}\right) = \frac{35}{128}$$

$$f(3) = N \cdot P(3) = 128\left(\frac{35}{128}\right) = 35$$

$$P(4) = \left(\frac{7-3}{3+1}\right)P(3)$$

$$P(4) = \frac{4}{4}P(3) = \left(\frac{35}{128}\right)$$

$$f(4) = N \cdot P(4) = 128\left(\frac{35}{128}\right) = 35$$

$$P(5) = \left(\frac{7-4}{4+1}\right)P(4)$$

$$P(5) = \frac{3}{5}p(4) = \left(\frac{3}{5}\right)\left(\frac{35}{128}\right) = \left(\frac{21}{128}\right)$$

$$f(5) = N \cdot P(5) = 128\left(\frac{21}{128}\right) = 21$$

$$P(6) = \left(\frac{7+5}{5+1}\right)P(5)$$

$$P(6) = \frac{2}{6}P(5) = \left(\frac{1}{3}\right)\left(\frac{21}{128}\right) = \left(\frac{7}{128}\right)$$

$$f(6) = N \cdot P(6) = 128\left(\frac{7}{128}\right) = 7$$

$$P(7) = \left(\frac{7-6}{6+1}\right)P(6)$$

$$P(7) = \frac{1}{7}P(6) = \left(\frac{1}{7}\right)\left(\frac{7}{128}\right) = \left(\frac{1}{128}\right)$$

$$f(7) = N \cdot P(7) = 128\left(\frac{1}{128}\right) = 1$$

The expected frequencies of the data are as follows:

Number of heads	0	1	2	3	4	5	6	7	Total
Frequency	7	6	19	35	30	23	7	1	128
Expected frequency, $f(x)$	1	7	21	35	35	21	1	1	128

(ii) The nature of the coin is not known:

Mean of the Binomial distribution = Mean of the distribution

$$np = \frac{\sum x_i f_i}{\sum f_i}$$

$$7p = \frac{0\times 7+1\times 6+2\times 19+3\times 35+4\times 30+5\times 23+6\times 7+7\times 1}{128}$$

$$= \frac{433}{128}$$

$$7p = 3.382813$$

$$P = 3.382813/7$$

$$= 0.483259$$

$$q = 1-p$$

$$= 1-0.483259$$

$$= 0.516741$$

Number of coins, $n = 7$

The recurrence relation for probabilities of Binomial distribution is

$$P(x+1) = \left(\frac{n-x}{x+1}\right)\frac{p}{q}P(x)$$

$$\text{Since } \frac{p}{q} = \frac{0.483259}{0.516741} = 0.935205$$

$$= P(x+1) = \left(\frac{n-x}{x+1}\right)\frac{p}{q}P(x)$$

The expected frequencies are calculated using,

$$f(x) = N \cdot P(x) \text{ where } N = 128$$

$$P(0) = 7C_0(0.483259)^0(0.516741)^7$$

$$= (0.516741)^7 = 0.009838$$

$$f(0) = N \cdot P(0)$$

$$= 128\ (0.009838)$$

$$= 1.259264 \sim 1$$

$$P(1) = \left(\frac{7-0}{0+1}\right)\frac{p}{q}P(0)$$

$$P(1) = 7\ (0.935205)\ P(0) = 7\ (0.935205)\ (0.009838)$$

$$= 0.064404$$

$$f(1) = N \cdot P(1) = 128(0.064404) = 8.243712 \sim 8$$

$$P(2) = \left(\frac{7-1}{1+1}\right)\frac{p}{q}P(1)$$

$$P(2) = \frac{6}{2}\ (0.0935205)\ P(1) = 3(0.935205)\ (0.0644404) = 0.180693$$

$$f(2) = N \cdot P(2) = 128(0.180693) = 23.1287 \sim 23$$

$$P(3) = \left(\frac{7-2}{2+1}\right)\frac{p}{q}P(2)$$

$$P(3) = \frac{5}{3}\frac{\mathrm{p}}{q}P(2) = \left(\frac{5}{3}\right)(0.935205)(0.180693) = 0.281642$$

$$f(3) = N \cdot P(3) = 128(0.281642) = 36.05018 \sim 36$$

$$P(4) = \left(\frac{7-3}{3+1}\right)\frac{p}{q}P(3)$$

$$P(4) = \frac{4}{4}\frac{\mathrm{p}}{q}P(3) = (0.935205)(0.281642) = 0.263393$$

$$f(4) = N \cdot P(4) = 128(0.263393) = 33.7143 \sim 34$$

$$P(5) = \left(\frac{7-4}{4+1}\right)\frac{p}{q}P(4)$$

$$P(5) = \frac{3}{5}\frac{\mathrm{p}}{q}P(4) = \left(\frac{3}{5}\right)(0.935205)(0.263393) = 0.147796$$

$$f(5) = N \cdot P(5) = 128(0.147796) = 18.91789 \sim 19$$

$$P(6) = \left(\frac{7-5}{5+1}\right)\frac{p}{q}P(5)$$

$$P(6) = \frac{2}{6}\frac{\mathrm{p}}{q}P(5) = \left(\frac{1}{3}\right)(0.935205)(0.147796) = 0.046073$$

$$f(6) = N \cdot P(6) = 128(0.046073) = 5.897344 \sim 6$$

$$P(7) = \left(\frac{7-6}{6+1}\right)\frac{p}{q}P(6)$$

$$P(7) = \frac{1}{7}\frac{\mathrm{p}}{q}P(6) = \left(\frac{1}{7}\right)(0.935205)(0.046073) = 0.006155$$

$$f(7) = N \cdot P(7) = 128(0.006155)$$
$$= 0.78784 \sim 1$$

The expected frequencies of the data are as follows:

Number of heads	0	1	2	3	4	5	6	7	Total
Frequency	7	6	19	35	30	23	7	1	128
Expected Frequency, $f(x)$	1	8	23	36	34	19	6	1	128

4.2 POISSON DISTRIBUTION

Definition

A random variable X is said to follow a Poisson distribution if it assumes only non-negative values and its probability distribution is given by,

$$P(X = x) = \begin{cases} \dfrac{e^{-\lambda}\lambda^x}{x!}, & x = 0, 1, 2, 3... \\ 0, & \text{otherwise} \end{cases}$$

Caution:

- Here, λ is called the parameter of the distribution.

  The above distribution is a probability mass function since

$$\Sigma_{x=0}^{\infty} P(X = x) = \Sigma_{x=0}^{\infty} \frac{e^{-\lambda}\lambda^x}{x!}$$
$$= e^{-\lambda} \Sigma_{x=0}^{\infty} \frac{\lambda^x}{x!}$$
$$= e^{-\lambda}\{1 + \frac{\lambda^1}{1!} + \frac{\lambda^2}{2!} + \frac{\lambda^3}{3!} + ...\}$$
$$= e^{-\lambda}.e^{\lambda}$$
$$= 1$$

$$\Sigma_{x=0}^{\infty} \frac{\lambda^x}{x!} \Sigma = \Sigma_{x=0}^{\infty} \frac{\lambda^{x-1}}{(x-1)!} = \Sigma_{x=0}^{\infty} \frac{\lambda^{x-2}}{(x-2)!} = e^{\lambda}$$

The following assumptions are made for Poisson distribution

- n, the number of trials is very large, that is, $n \to \infty$
- p, the constant probability of success for each trial is very small, that is, $p \to \infty$
- $np = \lambda$, is finite

The following are some instances, where Poisson distribution may be successfully employed:

- Number of printing mistakes at each page of a good book
- Number of air accidents in some unit of time
- Number of suicides reported in a particular city
- Number of defective material in a packing of a good concern

Moments of Poisson Distribution

Mean of Poisson Distribution

If X is a poisson variate, mean of X is given by,

$$E(X) = \sum_{x=0}^{\infty} x\, P(X = x)$$

$$= \sum_{x=0}^{\infty} x \frac{e^{-\lambda}\lambda^x}{x!}$$

$$= \sum_{x=0}^{\infty} \frac{e^{-\lambda}\lambda^x}{(x-1)!}$$

$$= e^{-\lambda} \sum_{x=0}^{\infty} \frac{\lambda^{x-1}}{(x-1)!}$$

$$E(X) = e^{-\lambda}\lambda \sum_{x=0}^{\infty} \frac{\lambda^{x-1}}{(x-1)!}$$

$$= e^{-\lambda}\lambda e^{\lambda}$$

$$E(X) = \lambda$$

Variance of Poisson Distribution

If X is a poisson variate, variance of X is given by,

$$V(X) = E(X^2) - \{E(X)\}^2$$

$$E(X^2) = \sum_{x=0}^{\infty} x^2 P(X = x)$$

$$= \sum_{x=0}^{\infty} x^2 \frac{e^{-\lambda}\lambda^x}{x!}$$

$$= \sum_{x=0}^{\infty} [x(x-1) + x] \frac{e^{-\lambda}\lambda^x}{x!}$$

$$= \sum_{x=0}^{\infty} x(x-1) \frac{e^{-\lambda}\lambda^x}{x!} + \sum_{x=0}^{\infty} x \frac{e^{-\lambda}\lambda^x}{x!}$$

$$= \sum_{x=0}^{\infty} x(x-1) \frac{e^{-\lambda}\lambda^x}{x!} + \lambda, \quad \{\text{since } E(X) = \lambda\}$$

$$= \sum_{x=0}^{\infty} \frac{e^{-\lambda}\lambda^2\lambda^{x-2}}{(x-2)!} + \lambda$$

$$= e^{-\lambda}\lambda^2 \sum_{x=0}^{\infty} \frac{\lambda^{x-2}}{(x-2)!} + \lambda$$

$$= e^{-\lambda}\lambda^2 e^{\lambda} + \lambda$$

$$= \lambda^2 + \lambda$$

$$E(X^2) = \lambda^2 + \lambda$$

$$V(X) = E(X^2) - \{E(X)\}^2$$

$$= \lambda^2 + \lambda - \lambda^2$$

$$= \lambda$$

Caution:

Mean of poisson distribution = variance of poisson distribution = λ

Worked Out Examples

EXAMPLE 4.13

If a poisson variate X is such that $P(x = 1) = 2\ P(x = 2)$, find the following:

(i) $P(x = 0)$

(ii) Mean

(iii) Variance of X

Solution: The probability distribution of Poisson distribution is given by,

$$P(X = x) = \frac{e^{-\lambda}\lambda^{x}}{x!}$$

Given that $P(x = 1) = 2\ P(x = 2)$

$$\frac{e^{-\lambda}\lambda^{1}}{1!} = 2\frac{e^{-\lambda}\lambda^{2}}{2!}$$

$$\frac{\lambda^{1}}{1!} = 2\frac{\lambda^{2}}{2!}$$

$$\lambda = 1$$

(i)
$$P(X = 0) = \frac{e^{-\lambda}\lambda^{0}}{0!}$$
$$= e^{-1}$$
$$= 0.367879$$

(ii) Mean of Poisson distribution = $\lambda = 1$

(iii) Variance of Poisson distribution = $\lambda = 1$

EXAMPLE 4.14

If X is a poisson variate such that $3P(x = 4) = \frac{1}{2}\ P(x = 2) + P(x = 0)$, find the following:

(i) Mean of X

(ii) $P(x \le 2)$

Solution: The probability distribution of Poisson distribution is given by,

$$P(X = x) = \frac{e^{-\lambda}\lambda^{x}}{x!}$$

Given that $3P(x = 4) = \frac{1}{2}\ P(x = 2) + P(x = 0)$,

$$3\frac{e^{-\lambda}\lambda^{4}}{4!} = \left(\frac{1}{2}\right)\frac{e^{-\lambda}\lambda^{2}}{2!} + \frac{e^{-\lambda}\lambda^{0}}{0!}$$

$$3\frac{\lambda^{4}}{24} = \frac{\lambda^{2}}{4} + 1$$

$$3\lambda^4 - 6\lambda^2 - 24 = 0$$
$$\lambda^4 - 2\lambda^2 - 8 = 0$$
$$(\lambda^2 - 4)(\lambda^2 + 2) = 0$$
$$\lambda^2 = 4$$
$$\lambda = +2, \text{ since } \lambda > 0$$

(i) Mean of Poisson distribution = $\lambda = +2, -2$

(ii) When $\lambda = +2$,

$$\begin{aligned} P(x \le 2) &= P(X = 0) + P(X = 1) + P(X = 2) \\ &= \frac{e^{-\lambda}\lambda^0}{0!} + \frac{e^{-\lambda}\lambda^1}{1!} + \frac{e^{-\lambda}\lambda^2}{2!} \\ &= e^{-\lambda}\{^0 + \lambda^0 + \lambda^1 + \lambda^2/2\} \\ &= e^{-2}\{1 + 2 + 4/2\} \\ &= 5e^{-2} \\ &= 0.676676 \end{aligned}$$

Fitting of Poisson Distribution

A probability distribution is to be fit for a given set of data points. Here, fitting of Poisson distribution is explained in detail. The main aim of fitting of a distribution is to find expected frequencies for a set of data points (x_i, f_i) which contain x_i and corresponding observed frequencies f_i. In order to find these, we find the probabilities of the Poisson distribution using the recurrence relation which is given in the next section. Using these probabilities, we find the expected frequencies $f(x) = NP(x)$, where $P(x)$ denotes the probabilities for $x = 0, 1, 2 \ldots n$. $N = \sum f_i$ where f_i denotes the observed frequencies of given data.

Recurrence Relation of Probabilities of Poisson Distribution

The probability distribution of Poisson distribution is

$$P(X = x) = P(x) = \frac{e^{-\lambda}\lambda^x}{x!} \tag{4.7}$$

$$P(x+1) = \frac{e^{-\lambda}\lambda^{x+1}}{(x+1)!} \tag{4.8}$$

Dividing equations (4.2) by (4.1) we get,

$$\begin{aligned} \frac{P(X+x)}{P(x)} &= \frac{\dfrac{e^{-\lambda}\lambda^{x+1}}{(x+1)!}}{\dfrac{e^{-\lambda}\lambda^x}{x!}} \\ &= \frac{\lambda^{x+1}x!}{\lambda^x(x+1)!} \\ &= \frac{\lambda}{(x+1)} \\ P(x+1) &= \frac{\lambda}{(x+1)}P(x) \end{aligned}$$

Worked Out Examples

EXAMPLE 4.15

Fit a Poisson distribution to the following data with respect to the number of red blood corpuscles (x) per cell:

x	0	1	2	3	4	5
Number of cells, f	142	156	69	27	5	1

Solution: Mean of the Poisson distribution = Mean of the above distribution

$$\lambda = \frac{\sum x_i fi}{\sum fi}$$

$$\lambda = \frac{0\times142+1\times156+2\times69+3\times27+4\times5+5\times1}{142+156+69+27+5+1}$$

$$= 1$$

Number of red corpuscles, $n = 5$

The recurrence relation for probabilities of Poisson distribution is

$$P(x+1) = \left(\frac{\lambda}{x+1}\right)P(x)$$

The expected frequencies are calculated using,

$f(x) = N \cdot P(x)$ where $N = 400$

$$P(0) = \frac{e^{-\lambda}\lambda^x}{x!} = \frac{e^{-\lambda}\lambda^0}{0!} = e^{-1}$$

$$= 0.367879$$

$$f(0) = N \cdot P(0)$$

$$= 400\ (0.367879)$$

$$= 147.1516 \sim 147$$

$$P(1) = \left(\frac{1}{0+1}\right)P(0)$$

$$P(1) = 0.367879$$

$$f(1) = N \cdot P(1) = 400\ (0.367879)$$

$$= 147.1516 \sim 147$$

$$P(2) = \left(\frac{1}{1+1}\right)P(1)$$

$$= \frac{1}{2}P(1)$$

$$= \frac{1}{2}(0.367879)$$

$$= 0.18394$$

$$f(2) = N \cdot P(2)$$

$$= 400(0.18394)$$

$$= 73.576 \sim 74$$

$$P(3) = \left(\frac{1}{2+1}\right)P(2)$$

$$P(3) = \frac{1}{3}P(2)$$

$$= \left(\frac{1}{3}\right)(0.18394)$$

$$= 0.061313$$

$$f(3) = N \cdot P(3)$$

$$= 400(0.61313)$$

$$= 24.5252 \sim 25$$

$$P(4) = \left(\frac{1}{3+1}\right)P(3)$$

$$P(4) = \frac{1}{4}P(3)$$

$$= \frac{1}{4}(0.061313)$$

$$= 0.015328$$

$$f(4) = N \cdot P(4)$$

$$= 400(0.015328)$$

$$= 6.1312 \sim 6$$

$$P(5) = \left(\frac{1}{4+1}\right)P(4)$$

$$P(5) = \frac{1}{5}P(4) = \left(\frac{1}{5}\right)(0.015328)$$

$$= 0.003066$$

$$f(5) = N \cdot P(5) = 400(0.003066)$$

$$= 1.2264 \sim 1$$

The expected frequencies of the given data are obtained as follows:

x	0	1	2	3	4	5	Total
Number of cells, f	142	156	69	27	5	1	400
Expected frequency, $f(x)$	147	147	74	25	6	1	400

EXAMPLE 4.16

Fit a Poisson distribution to the following data by calculating the expected frequencies:

x	0	1	2	3	4	5	6	7	8
f	71	112	117	57	27	11	3	1	1

Solution: Mean of the Poisson distribution = Mean of the above distribution

$$\lambda = \frac{\sum x_i fi}{\sum fi}$$

$$\lambda = \frac{0\times 71+1\times 112+2\times 117+3\times 57+4\times 27+5\times 11+6\times 3+7\times 1+8\times 1}{71+112+117+57+27+11+3+1+1}$$

$$= \frac{713}{400}$$

$$= 1.7825$$

$$n = 8$$

The recurrence relation for probabilities of Poisson distribution is

$$P(x+1) = \left(\frac{\lambda}{x+1}\right)P(x)$$

The expected frequencies are calculated using
$f(x) = N \cdot P(x)$, where $N = 400$

$$P(0) = \frac{e^{-\lambda}\lambda^x}{x!} = \frac{e^{-\lambda}\lambda^0}{0!} = e^{-1.7825}$$

$$= 0.168217$$

$$f(0) = N \cdot P(0)$$

$$= 400\ (0.168217)$$

$$= 67.2868 \sim 67$$

$$P(1) = \left(\frac{1.7825}{0+1}\right)P(0)$$

$$P(1) = (1.7825)(0.168217)$$

$$= 0.299847$$

$$f(1) = N \cdot P(1) = 400\ (0.299847)$$
$$= 119.9388 \sim 120$$
$$P(2) = \left(\frac{1.7825}{1+1}\right)P(1)$$
$$= \left(\frac{1.7825}{2}\right)P(1)$$
$$= \left(\frac{1.7825}{2}\right)(0.299847) = 0.267239$$
$$f(2) = N \cdot P(2)$$
$$= 400(0.267239)$$
$$= 106.8956 \sim 107$$
$$P(3) = \left(\frac{1.7825}{2+1}\right)P(2)$$
$$P(3) = \frac{1.7825}{3}P(2)$$
$$= \left(\frac{1.7825}{3}\right)(0.267239)$$
$$= 0.158785$$
$$f(3) = N \cdot P(3)$$
$$= 400(0.158785)$$
$$= 63.514 \sim 64$$
$$P(4) = \left(\frac{1.7825}{3+1}\right)P(3)$$
$$P(4) = \frac{1.7825}{4}P(3)$$
$$= \frac{1.7825}{4}(0.158785)$$
$$= 0.070759$$
$$f(4) = N \cdot P(4)$$
$$= 400(0.070759)$$
$$= 28.3036 \sim 28$$
$$P(5) = \left(\frac{1.7825}{4+1}\right)P(4)$$
$$P(5) = \frac{1.7825}{5}P(4) = \left(\frac{1.7825}{5}\right)(0.070759)$$
$$= 0.025226$$

$$f(5) = N \cdot P(5) = 400(0.025226)$$
$$= 10.0904 \sim 10$$
$$P(6) = \left(\frac{1.7825}{5+1}\right)P(5)$$
$$P(6) = \frac{1.7825}{5+1}P(5) = \left(\frac{1.7825}{6}\right)(0.025226)$$
$$= 0.007494$$
$$f(6) = N \cdot P(6) = 400(0.007494)$$
$$= 2.9976 \sim 3$$
$$P(7) = \left(\frac{1.7825}{6+1}\right)P(6)$$
$$P(7) = \frac{1.7825}{7}P(6) = \left(\frac{1.7825}{7}\right)(0.007494)$$
$$= 0.001908$$
$$f(7) = N \cdot P(7) = 400(0.001908)$$
$$= 0.7632 \sim 1$$
$$P(8) = \left(\frac{1.7825}{7+1}\right)P(7)$$
$$P(8) = \frac{1.7825}{8}P(7) = \left(\frac{1.7825}{8}\right)(0.001908)$$
$$= 0.000425$$
$$f(8) = N \cdot P(8) = 400(0.000425)$$
$$= 0.17 \sim 0$$

The expected frequencies of the given data are obtained as follows:

x	0	1	2	3	4	5	6	7	8	Total
f	71	112	117	57	27	11	3	1	1	400
$f(x)$	67	120	107	64	28	10	3	1	0	400

EXAMPLE 4.17

After correcting 50 pages of the proof of a book, the proofreader finds that there are on the average 2 errors per 5 pages. How many pages would one expect to find with 0, 1, 2, 3, and 4 errors in 1,000 pages of the first print of the book?

Solution: Let X denotes the number of errors per page.

Then the mean number of errors per page is given by

$$\lambda = \frac{2}{5} = 0.4$$

Using Poisson distribution,
Probability that there are x errors per page $= P(X = x)$

$$P(X = x) = P(x) = \frac{e^{-\lambda}\lambda^x}{x!}, \quad X = 0, 1, 2, 3, \ldots$$

The expected number of pages with x errors per page in a book of 1,000 pages is

$$f(x) = 1000x\ P(x)$$

The recurrence relation for probabilities of Poisson distribution is

$$P(x+1) = \frac{\lambda}{x+1}P(x)$$

$$P(0) = e^{-\lambda}$$

$$= e^{-0.4}$$

$$= 0.6703$$

$$f(0) = 1000\ (0.6703) = 670.3 \approx 670$$

$$P(1) = \frac{\lambda}{0+1}P(0)$$

$$= (0.4)(0.6703)$$

$$= 0.26812$$

$$f(1) = 1000\ (0.26812)$$

$$= 268.12 \approx 268$$

$$P(2) = \frac{\lambda}{1+1}P(1)$$

$$= \frac{0.4}{2}(0.26812)$$

$$= 0.053624$$

$$f(2) = 1000\ (0.053624)$$

$$= 53.624 \approx 54$$

$$P(3) = \frac{\lambda}{2+1}P(2)$$

$$= \frac{0.4}{3}(0.053624)$$

$$= 0.0071298$$

$$f(3) = 1000\ (0.0071298)$$

$$= 7.1298 \approx 7$$

$$P(4) = \frac{\lambda}{3+1}P(3)$$

$$= \frac{0.4}{4}(0.0071298)$$

$$f(4) = 1000\ (0.00071298)$$
$$= 0.71928 \approx 1$$

The expected number of pages with x errors is obtained as

No. of errors per page, x	0	1	2	3	4
Expected number of pages, $f(x)$	670	268	54	7	1

EXAMPLE 4.18

Fit a Poisson distribution for the following data and calculate the expected frequencies:

x	0	1	2	3	4
f	109	65	22	3	1

Solution: Here, mean of the Poisson distribution = Mean of the given distribution

$$\lambda = \frac{\sum x_i f_i}{\sum f_i}$$
$$N = \sum f_i = 109 + 65 + 22 + 3 + 1 = 200$$
$$\lambda = \frac{0 + 65 + 44 + 9 + 4}{200}$$
$$= \frac{122}{200}$$
$$\lambda = 0.61$$

Using the recurrence relation for probabilities, we can find $P(x)$ using,

$$P(x+1) = \frac{\lambda}{x+1} P(x)$$

The expected frequencies are obtained by,

$$f(x) = N \cdot P(x)$$

The first probability is obtained using,

$$P(0) = e^{-\lambda} = e^{-0.61} = 0.543351$$
$$f(0) = N \cdot P(0) = 200(0.543351)$$
$$= 108.6702 \sim 109$$
$$P(1) = \frac{\lambda}{0+1} P(0)$$
$$= \frac{0.61}{1}(0.543351)$$
$$= 0.331444$$

$$f(1) = N \cdot P(1) = 200(0.331444)$$
$$= 66.2888 \sim 66$$
$$P(2) = \frac{\lambda}{1+1} P(1)$$
$$= \frac{0.61}{2}(0.331444)$$
$$= 0.10109$$
$$f(2) = N \cdot P(2) = 200(0.10109)$$
$$= 20.218 \sim 20$$
$$P(3) = \frac{\lambda}{2+1} P(2)$$
$$= \frac{0.61}{3}(0.10109)$$
$$= 0.20555$$
$$f(3) = N \cdot P(3) = 200(0.020555)$$
$$= 4.111 \sim 4$$
$$P(4) = \frac{\lambda}{3+1} P(3)$$
$$= \frac{0.61}{4}(0.020555)$$
$$= 0.003135$$
$$f(4) = N \cdot P(4) = 200(0.003135)$$
$$= 0.627 \sim 1$$

Hence, the expected frequencies are:

x	0	1	2	3	4	Total
Observed frequencies, f	109	65	22	3	1	200
Expected frequencies, $f(x)$	109	66	20	4	1	200

The MGF of Poisson Distribution

Let X be a discrete random variable which follows Binomial distribution. Then

$$M_x(t) = E(e^{tx}) = \sum_{x=0}^{\infty} e^{tx} \frac{e^{-\lambda} \lambda^x}{x!}$$
$$= \sum_{x=0}^{\infty} \frac{e^{-\lambda} (\lambda e^t)^x}{x!}$$
$$= e^{-\lambda} \{1 + \lambda e^t + \frac{(\lambda e^t)^2}{2!} + \cdots\}$$
$$= e^{-\lambda} e^{\lambda e^t}$$
$$= e^{\lambda(e^t - 1)}$$

The MGF About Mean of Poisson Distributions

Since the mean of Poisson distribution is λ, we have

$$\begin{aligned} E\{e^{t(x-\lambda)}\} &= E\{e^{tx}e^{-t\lambda}\} \\ &= e^{-t\lambda}E(e^{tx}) \\ &= e^{-t\lambda}M_x(t) \\ &= e^{-t\lambda}e^{\lambda(e^t-1)} \\ &= e^{-\lambda}e^{\lambda(e^t-1)} \end{aligned}$$

Additive Property of Poisson Distribution

Let X and Y be two independent Poisson variables with λ_1 and λ_2 as the parameters.

Then the MGF's of X and Y are given by

$$\begin{aligned} M_x(t) &= e^{\lambda 1(e^t-1)} \\ M_y(t) &= e^{\lambda 2(e^t-1)} \\ M_{x+y}(t) &= M_x(t)\cdot M_y(t), \text{ since } X \text{ and } Y \text{ are independent.} \\ &= e^{\lambda 1(e^t-1)} \ .e^{\lambda 2(e^t-1)} \\ M_{x+y}(t) &= e^{(\lambda 1+\lambda 2)(e^t-1)} \end{aligned}$$

This is the MGF of a Poisson variable with parameter $(\lambda_1 + \lambda_2)$. Hence by Uniqueness theorem of MGF's, $X + Y$ is also a Poisson variate with parameter $(\lambda_1 + \lambda_2)$.

Characteristic Function of Poisson Distribution

$$\begin{aligned} \varphi_{x(t)} &= E(e^{itx}) \\ &= \sum_{x=0}^{\infty} e^{itx}\frac{e^{-\lambda}\lambda^x}{x!} \\ &= \sum_{x=0}^{\infty}\frac{e^{-\lambda}(\lambda e^{it})^x}{x!} \\ &= e^{-\lambda}\sum_{x=0}^{\infty}\frac{(\lambda e^{it})^x}{x!} \\ &= e^{-\lambda}e^{\lambda eit} \\ &= e^{\lambda(e^{it}-1)} \end{aligned}$$

4.3 NEGATIVE BINOMIAL DISTRIBUTION

Suppose we have a succession of n Bernoulli trials. We assume that

(i) The trials are independent

(ii) The probability of success p is constant from trial to trial.

Let $f(x)$ denote the probability that there are x failures preceding the r^{th} success in $x + r$ trials. The last trial is a success, whose probability is p. In the remaining $(x + r - 1)$ trials we have $(r - 1)$ successes. This probability is given by the Binomial distribution by

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

Hence, by compound probability theorem the probability is given by the product of these two probabilities as

$$[(x+r+1)C_{r-1}p^{r-1}q^x][p]$$

Definition

A random variable X is said to follow negative Binomial distribution if its probability mass function

$$P(X=x)=\begin{cases}(x+r-1)C_{r-1}p^r q^x, & x=0,1,2,...\\ 0, & \text{otherwise}\end{cases} \tag{4.9}$$

Caution:

$$\text{Since } nC_r = nC_{n-r}, (x+r+1)C_{r-1} = (x+r-1)C_x$$
$$= (-1)^x(-r)C_x$$

$$P(X=x)=\begin{cases}(-r)C_x p^r(-q)^x, & x=0,1,2,...\\ 0, & \text{otherwise}\end{cases} \tag{4.10}$$

$$\sum_{x=0}^{\infty}P(X=x) = p^r\sum_{x=0}^{\infty}(-r)C_x(-q)^x$$
$$= p^r(1-q)^{-r}$$
$$= 1$$

Let $p=\frac{1}{Q}$ and $q=\frac{P}{Q}$ such that $Q-P=1$, then

$$P(X=x)=\begin{cases}(-r)C_x Q^{-r}\left(-\frac{P}{Q}\right)^x, & x=0,1,2,...\\ 0, & \text{otherwise}\end{cases} \tag{4.11}$$

The above term gives the general term of the binomial expansion $(Q-P)^{-r}$.

The MGF of Negative Binomial Distribution

Let X be a discrete random variable which follows negative Binomial distribution. Then the MGF of X is given by,

$$M_x(t) = E(e^{tx})$$
$$=\sum_{x=0}^{\infty}e^{tx}P(X=x)$$
$$=\sum_{x=0}^{\infty}e^{tx} - rC_x Q^{-r}\left(-\frac{P}{Q}\right)x$$
$$=\sum_{x=0}^{\infty}-rC_x Q^{-r}\left(-\frac{Pe^t}{Q}\right)x$$
$$= (Q-Pe^t)^{-r}$$

Moments of Negative Binomial Distribution

The mean of negative Binomial distribution is obtained from MGF as

$$\mu_1^1 = \left[\frac{d}{dt}M_x(t)\right]t=0$$
$$= [-r(Q-Pe^t)^{-r-1}(-Pe^t)]_{t=0}$$

$$= (-r)(-p)$$
$$= rP$$

Hence, the mean of negative Binomial distribution is *rP*.

$$\mu_2^1 = \left[\frac{d^2}{dt^2} M_x(t)\right] t = 0$$
$$= [-r(Q - Pe^t)^{-r-1}(-Pe^t)] + [(-r)(-Pe^t)(-r-1)(Q - Pe^t)^{-r-2}(-Pe^t)]_{t=0}$$
$$= rP + r(r+1)P^2$$

$$\mu_2 = \mu_2^1 - [\mu_1^1]^2$$
$$= rP + r(r+1)P^2 - (rP)^2$$
$$= rP + r^2P^2 + rP^2 - r^2P^2$$
$$= rP + rP^2$$
$$\mu_2 = rP(1+P)$$
$$= rPQ$$

Hence, the variance of negative Binomial distribution is *rPQ.*
Caution: The mean of negative Binomial distribution is *rP* and variance is *rPQ.*

4.4 GEOMETRIC DISTRIBUTION

Suppose we have a series of independent trials or repetitions and on each repetition the probability of success remains the same. Then the probability that there are *x* failures preceding the first success is $q^x p$.

Definition

A random variable *X* is said to follow geometric distribution if it assumes only non-negative values and its probability mass function is given by

$$P(X = x) = \begin{cases} q^x p, & x = 0, 1, 2, 3, \ldots 0 < p < 1 \\ 0, & \text{otherwise} \end{cases} \tag{4.12}$$

Caution:

- Since the various probabilities for $x = 0, 1, 2, \ldots$ are the various terms of geometric progression, hence the name geometric distribution.
- The above distribution shows probability mass function since

$$\sum_{x=0}^{\infty} P(X = x) = \sum_{x=0}^{\infty} q^x p$$
$$= p + pq + pq^2 + \cdots$$
$$= p(1 + q + q^2 + \cdots)$$
$$= \frac{p}{1-q}$$
$$= 1$$

Moments of Geometric Distribution

$$\mu_1^1 = \sum_{x=1}^{\infty} x\,P(X=x) = \sum_{x=1}^{\infty} x q^x p$$
$$= pq\sum_{x=1}^{\infty} x q^{x-1}$$
$$= pq\{1+2q+3q^2+\cdots\}$$
$$= pq(1-q)^{-2}$$
$$= \frac{q}{p}$$
$$\mu_1^1 = E(X) = \frac{q}{p}$$
$$\mu_2^1 = E(X^2) = \sum_{x=1}^{\infty} x^2 P(X=x)$$
$$\mu_2^1 = \sum_{x=2}^{\infty} [x(x-1)+x]q^x p$$
$$= \sum_{x=2}^{\infty} [x(x-1)]q^x p + \sum_{x=1}^{\infty} x q^x p$$
$$= 2\sum_{x=2}^{\infty} \frac{x(x-1)}{2.1} q^{x-2} q^2 p + \frac{q}{p}$$
$$= 2p\,q^2(1-q)^{-3} + \frac{q}{p}$$
$$\therefore \mu_2^1 = 2p^{-2}q^2 + \frac{q}{p}$$

Variance of X is given by,

$$V(X) = E(X^2) - [E(X)]^2$$
$$= 2\frac{q^2}{p^2} + \frac{q}{p} - \frac{q^2}{p^2}$$
$$= \frac{q}{p} + \frac{q^2}{p^2}$$
$$= \frac{pq+q^2}{p^2}$$
$$= \frac{q(p+q)}{p^2}$$
$$= \frac{q}{p^2}$$

The MGF of Geometric Distribution

Let X be a discrete random variable which follows geometric distribution. Then the MGF of X is given by

$$M_x(t) = E(e^{tx})$$
$$= \sum_{x=0}^{\infty} e^{tx} q^x p$$
$$= p\sum_{x=0}^{\infty} e^{tx} q^x$$
$$= p\sum_{x=0}^{\infty} (e^t q)^x$$
$$= p[1+(e^t q)^1 + (e^t q)^2 + (e^t q)^3 + \cdots]$$

$$= p[1-(e^t q)]^{-1}$$

$$= \frac{p}{1-qe^t}$$

$$M_x(t) = \frac{p}{1-qe^t}$$

4.5 HYPER GEOMETRIC DISTRIBUTION

This is a discrete distribution and is defined as follows:

Definition

A discrete random variable X is said to follow the hyper geometric distribution if it assumes only non-negative values and its probability mass function is given by,

$$P(X = x) = \begin{cases} \dfrac{(MC_K)(N - MC_{n-K})}{NC_n}, & k = 0, 1, 2, 3, \ldots \min(n, M) \\ 0, & \text{otherwise} \end{cases} \tag{4.13}$$

Caution:

- The total number of samples of size n chosen from N items is NC_n which are equally likely. There are MC_K ways of selecting K successes from the M that are available, and for each of these ways we can choose $n - \kappa$ failures in $N - MC_{n-\kappa}$ ways.
- Hence the total number of favourable samples among NC_n possible samples is given by (MC_κ) $(N - MC_{n-\kappa})$.
- This distribution is assumed when the population is finite and sampling is done without replacement.
- The events here are dependent although random.
- N, M, and n are parameters of this distribution.
- N is a positive integer, M is a positive integer which should not exceed N and n is a positive integer that is at most N.
- If an urn with N balls, M of which are white and $N - M$ which are red are considered and suppose a sample of n balls are drawn without replacement from the urn. Then the probability of getting k white balls out $n(k < n)$ is $\dfrac{(MC_K)(N - MC_{n-K})}{NC_n}$.

Moments of Hyper Geometric Distribution

Mean of Hyper Geometric Distribution

$$E(X) = \sum_{k=0}^{n} kP(X = k)$$

$$= \sum_{k=0}^{n} \left\{ k \frac{(MC_K)(N - MC_{n-K})}{N_{C_n}} \right\}$$

$$= \sum_{k=0}^{n} k \left\{ \frac{M}{k} \frac{(M - 1C_{K-1})(N - MC_{n-K})}{NC_n} \right.$$

$$= \frac{M}{NC_n} \sum_{k=1}^{n} (M-1C_{K-1})(N-MC_{n-K})$$

Let $x = k-1$, $m = n-1$, $M-1 = R$ (4.14)

$$E(X) = \frac{M}{NC_n} \sum_{x=0}^{m} (RC_x)(N-R-1C_{m-x})$$

$$= \frac{M}{NC_n} \left\{ (RC_0)(N-R-1C_{m-0}) + (RC_1)(N-R-1C_{m-1}) + \ldots (RC_n)(N-R-1C_{m-m}) \right\}$$

$$= \frac{M}{NC_n} (N-1C_m)$$

$$= \frac{M}{NC_n} (N-1C_{n-1}) \qquad \{\text{from eqn.(4.14)}\}$$

$$E(X) = \frac{nM}{N}$$

$$(X^2) = X(X-1) + X$$

$$E(X^2) = E\{X(X-1) + X\}$$

$$= E\{X(X-1)\} + E(X)$$

$$E\{X(X-1)\} = \sum_{k=0}^{n} k(k-1) \{ \frac{(MC_K)(N-MC_{n-k})}{NC_n}$$

$$= \frac{M(M-1)}{NC_n} \sum_{k=2}^{n} (M-2C_{k-2})(N-MC_{n-k})$$

$$= \frac{M(M-1)}{NC_n} (N-2C_{n-2})$$

$$= \frac{n(n-1)M(M-1)}{N(N-1)}$$

$$E(X^2) = \frac{n(n-1)M(M-1)}{N(N-1)} + \frac{nM}{N}$$

Variance of Hyper Geometric Distribution

$$V(X) = E(X^2) + \{E(X)\}^2$$

$$V(X) = \frac{n(n-1)M(M-1)}{N(N-1)} + \frac{nM}{N} - \left\{ \frac{nM}{N} \right\}^2$$

$$= \frac{NM(N-M)(N-n)}{N^2(N-1)}$$

4.6 UNIFORM DISTRIBUTION

This is a discrete distribution and is defined as follows:

Definition

A random variable X is said to have a discrete uniform distribution if its probability mass function is defined by

$$P(X = x) = \begin{cases} \frac{1}{n}, & \text{for } x = 1, 2, \ldots n \\ 0, & \text{otherwise} \end{cases} \tag{4.15}$$

Caution:

- This is one where the random variable assumes each of its values with equal probability.
- Here, n is the parameter of the distribution.
- n takes only set of positive integers,
- When the different values of the random variable become equally likely, the distribution can be used.
- The distribution is appropriate for a die experiment and an experiment with a deck of cards.

Moments of Uniform Distribution

Mean of uniform distribution is given by,

$$\begin{aligned} E(X) &= \sum_{x=1}^{n} xP(X = x) \\ &= \sum_{x=1}^{n} x\left(\frac{1}{n}\right) \\ E(X) &= \left(\frac{1}{n}\right)\sum_{x=1}^{n} x \\ &= \left(\frac{1}{n}\right)\left(\frac{n(n+1)}{2}\right) \\ &= \left(\frac{n+1}{2}\right) \\ E(X^2) &= \sum_{x=1}^{n} x^2 P(X = x) \\ &= \sum_{x=1}^{n} x^2\left(\frac{1}{n}\right) \\ &= \frac{1}{n}\sum_{x=1}^{n} x^2 \\ &= \left(\frac{1}{n}\right)\left(\frac{n(n+1)(2n+1)}{6}\right) \\ &= \frac{(n+1)(2n+1)}{6} \end{aligned}$$

The variance of uniform distribution is given by,

$$\begin{aligned} V(X) &= E(X^2) - [E(X)]^2 \\ &= \frac{(n+1)(2n+1)}{6} - \left[\left(\frac{n+1}{2}\right)\right]^2 \\ &= \frac{(n+1)(2n+1)}{6} - \frac{(n+1)^2}{4} \\ &= \frac{2(n+1)(2n+1) - 3(n+1)^2}{12} \end{aligned}$$

$$= \frac{(n+1)(4n+2-3n-3)}{12}$$

$$V(X) = \frac{(n+1)(n-1)}{12}$$

The MGF of Uniform Distribution

Let X be a discrete random variable, which follows uniform distribution. Then the MGF of X is given by

$$M_x(t) = E(e^{tx})$$
$$= \sum_{x=1}^{n} e^{tx} P(X = x)$$
$$= \sum_{x=1}^{n} e^{tx} \left(\frac{1}{n}\right)$$
$$= \frac{1}{n} \sum_{x=1}^{n} e^{tx}$$
$$= \frac{1}{n}[e^t + e^{2t} + e^{3t} + \ldots e^{nt}]$$
$$= \frac{1}{n}[e^t + (e^t)^2 + (e^t)^3 + \ldots (et)^n]$$
$$= \frac{1}{n} e^t [1 + (e^t) + (e^t)^2 + \ldots (e^t)^{n-1}]$$
$$M_x(t) = \frac{e^t(1-e^{nt})}{n(1-e^t)}$$

DEFINITIONS AT A GLANCE

Binomial Distribution: A random variable X has a Binomial distribution if X assumes only non-negative values and its probability distribution is given by,

$$P(X = r) = \begin{cases} nC_r p^r q^{n-1}, r = 0, 1, \ldots n \\ 0, \qquad \text{otherwise} \end{cases}$$

Poisson Distribution: A random variable X is said to follow a Poisson distribution if it assumes only non-negative values and its probability distribution is given by,

$$P(X = x) = \begin{cases} \dfrac{e^{-\lambda}\lambda^x}{x!}, \ -x = 0, 1, 2, 3\ldots \\ 0, \qquad \text{otherwise} \end{cases}$$

Negative Binomial Distribution: A random variable X is said to follow negative Binomial distribution if its probability mass function

$$P(X = x) = \begin{cases} (x+r-1)C_{r-1} p^r q^x, \ x = 0, 1, 2,\ldots \\ 0, \qquad \text{otherwise} \end{cases}$$

Geometric Distribution: A random variable X is said to follow geometric distribution if it assumes only non-negative values and its probability mass function is given by,

$$P(X = x) = \begin{cases} q^x p, & x = 0, 1, 2, 3, \ldots \; 0 < p < 1 \\ 0, & \text{otherwise} \end{cases}$$

Hyper Geometric Distribution: A discrete random variable X is said to follow the hyper geometric distribution if it assumes only non-negative values and its probability mass function is given by,

$$P(X = x) = \begin{cases} \dfrac{(MC_k)(N - MC_{n-k})}{NC_n}, & k = 0, 1, 2, 3, \ldots \min(n, M) \\ 0, & \text{otherwise} \end{cases}$$

Uniform Distribution: A random variable X is said to have a discrete uniform distribution, if its probability mass function is defined by,

$$P(X = x) = \begin{cases} \dfrac{1}{n}, \text{ for } x = 1, 2, \ldots n \\ 0, \text{ otherwise} \end{cases}$$

FORMULAE AT A GLANCE

- Mean of Binomial distribution, $E(X) = np$
- Variance of Binomial distribution, $V(X) = npq$
- The recurrence relation for the probabilities of Binomial distribution is

$$P(x+1) = \left(\frac{n-x}{x+1}\right)\frac{p}{q}P(x)$$

- The MGF of Binomial distribution is

$$M_x(t) = (q + pe^t)^n$$

- The characteristic function of Binomial distribution is

$$\varphi_x(t) = (q + pe^{it})^n$$

- Mean of Poisson distribution = variance = λ
- The recurrence relation for the probabilities of Poisson distribution is

$$P(x+1) = \frac{\lambda}{x+1}P(x)$$

- The MGF of Poisson distribution is

$$M_x(t) = e^{\lambda(e^t - 1)}$$

- The characteristic function of Poisson distribution is

$$\varphi_x(t) = e^{\lambda} e^{(it-1)}$$

- The MGF of Negative Binomial distribution is

$$M_x(t) = (Q - Pe^t)^{-r}$$

- Mean of negative Binomial distribution is rP and variance is rPQ
- Mean of geometric distribution is $\frac{q}{p}$ and variance is $\frac{q}{p^2}$
- The MGF of geometric distribution is

$$M_x(t) = \frac{p}{1-q^{et}}$$

- Mean of hyper geometric distribution is $E(X) = \frac{nM}{N}$
- Mean of uniform distribution is $E(X) = \frac{n+1}{2}$ and variance = $\frac{(n+1)(n+2)}{12}$

OBJECTIVE TYPE QUESTIONS

1. The probability of getting at least one head when a coin is tossed 6 times is _____.

   (a) $\frac{63}{64}$ (b) $\frac{1}{64}$

   (c) $\frac{61}{64}$ (d) $\frac{59}{64}$

2. The mean and standard deviation of Binomial distribution are _____ and _____.

   (a) p, pq (b) np, np

   (c) $np, \sqrt{npq}$ (d) np, npq

3. The mean of Binomial distribution is 6 and variance is 2, then p = _____

   (a) $\frac{2}{3}$ (b) $\frac{1}{3}$

   (c) 1 (d) 0

4. A coin is tossed thrice. The probability of getting two heads is _____

   (a) $\frac{3}{8}$ (b) $\frac{5}{8}$

   (c) $\frac{1}{8}$ (d) $\frac{7}{8}$

5. If X is Poisson variate with $P(X=3) = \frac{1}{6}$ and $P(X=2) = \frac{1}{3}$, then $P(X=0)$ = _____.

   (a) $e^{-\frac{1}{2}}$ (b) $e^{-\frac{1}{2}}$

   (c) $e^{\frac{1}{2}}$ (d) $e^{\frac{3}{2}}$

6. The probability distribution in which mean = variance is _____.

   (a) Poisson (b) Binomial

   (c) Geometric (d) Uniform

7. If $P(1) = P(2)$, then the mean of Poisson distribution is _____.
   (a) 1 (b) 0
   (c) 2 (d) 3

8. If the standard deviation of Poisson distribution is m, then its mean = _____.
   (a) m (b) m^2
   (c) $\sqrt{m}$ (d) m

9. Mean of negative Binomial distribution is _____.
   (a) rP (b) rQ
   (c) P (d) rPQ

10. Mean of geometric distribution is _____ and variance is_____.
   (a) $\frac{q}{p}, \frac{q}{p}$ (b) $\frac{p}{q}, \frac{p}{q}$
   (c) $\frac{q}{p}, \frac{q}{p^2}$ (d) $\frac{p}{q}, \frac{p}{q^2}$

11. Mean of hyper geometric distribution is $E(X)$ = _____.
   (a) $\frac{nM}{N}$ (b) $\frac{M}{n}$
   (c) $\frac{N}{nM}$ (d) $\frac{M}{nN}$

ANSWERS

1. (a) 2. (c) 3. (a) 4. (a) 5. (b) 6. (a) 7. (c) 8. (c)
9. (a) 10. (c) 11. (a)

5 Standard Continuous Distributions

Prerequisites

Before you start reading this unit, you should:

- Know the concepts of mean and standard deviations and their calculations
- Know the definition of probability and the methods for calculating it

Learning Objectives

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

- To identify problems that can be modelled using normal distributions and know about other standard continuous distributions such as Exponential, Gamma, and Weibull distributions
- Know about central limit theorem and where it is applied
- Study all the chief characteristics of normal distribution
- To know the law of large numbers

INTRODUCTION

In the earlier units, we have gone through some discrete distributions. Now, let us go through some continuous distributions. Of all continuous distributions, normal distribution is very important. We shall study some of its characteristics, importance, some applications of this distribution.

5.1 NORMAL DISTRIBUTION

This is a continuous distribution.

Definition

A continuous random variable X is said to follow normal distribution if

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

$$-\infty < \mu < \infty$$

$$\sigma > 0$$

Caution:

- μ(called mean) and σ^2 (called variance) are called the parameters.
- $X \sim N(\mu, \sigma^2)$ is read as X follows normal distribution with parameters μ and σ^2.
- If $X \sim N(\mu, \sigma^2)$, then $Z = \frac{X - \mu}{\sigma}$ is a standard normal variate with

$$E(z) = E\left(\frac{X-\mu}{\sigma}\right) = \frac{1}{\sigma}[E(X) - E(\mu)]$$

$$= \frac{1}{\sigma}[\mu - \mu] = 0$$

$$V(Z) = V\left(\frac{X-\mu}{\sigma}\right) = V\left(\frac{X}{\sigma} - \frac{\mu}{\sigma}\right) = \frac{1}{\sigma^2}V(X)$$

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

$\therefore Z \sim N(0,1)$ is read as Z follows standard normal distribution with parameters 0 and 1.

- Some of the situations where normal distribution occurs are physical measurements in areas such as rainfall studies, meteorological experiments, errors in scientific measurements.
- The probability density function (pdf) of standard normal distribution is

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}, \quad -\infty < z < \infty$$

$$= \int_{-\infty}^{\infty} dx\,dx \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}} dx \qquad Z = \frac{x-\mu}{\sigma} \quad dz = \frac{1}{\sigma}dx$$

$$= \frac{1}{\cancel{\sigma}\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{z^2}{2}} \cancel{\sigma} dz = 1$$

Normal Distribution

Continuous Random variables are used in situations when we deal with quantities that are measured on a continuous scale, like the speed of a car.

Definition

If X is a continuous random variable, then the pdf is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \qquad -\infty < x < \infty$$

Chief Characteristics of Normal Distribution

1. The graph of Normal distribution which is the normal curve is bell-shaped.
2. The normal curve is symmetric about a vertical line $x = \mu$.
3. The mean, median, and mode of normal distribution coincide.
4. The normal curve has its point of inflection at $x = \mu \pm \sigma$, is concave downward if $\mu - \sigma < X < \mu + \sigma$ and is concave upward otherwise.
5. x-axis is an asymptote to the normal curve.
6. Since $f(x)$ being the probability can never be negative, no portion of the curve lies below the x-axis.
7. The maximum probability occurring at the point $X = \mu$ is given by $[P(x)]_{\max} = \frac{1}{\sigma\sqrt{2\pi}}$.
8. The odd ordered moments of normal distribution vanish and even ordered moments are given by $\mu_{2r} = 1 \cdot 3 \cdot 5 \cdots (2_{r-1})\sigma^{2r}$, $r = 0, 1, 2, \ldots$

9. Mean deviation about mean μ is $\sqrt{\frac{2}{\pi}}\sigma \simeq \frac{4}{5}\sigma$ (approximately).
10. Since the total area under the normal is 1, the area of the normal curve from $x = -\infty$ to $x = \mu$ is $\frac{1}{2}$ and from $x = \mu$ to $x = \infty$ is $\frac{1}{2}$.
11. $P(\mu - \sigma < X < \mu + \sigma) = 0.6826$
    $P(\mu - 2\sigma < X < \mu + 2\sigma) = 0.9544$
    $P(\mu - 3\sigma < X < \mu + 3\sigma) = 0.9973$

Mean of Normal Distribution

If X is a normal variate, then

$$E(X) = \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty} xe^{-\frac{1}{2}\left(x-\frac{\mu}{\sigma}\right)^2} dx$$

Let $z = \frac{x-\mu}{\sigma}, dz = \frac{1}{\sigma}dx, x = \mu + \sigma z$

$$E(X) = \frac{1}{\cancel{\sigma}\sqrt{2\pi}}\int_{-\infty}^{\infty} (\mu + \sigma z)e^{\frac{-z^2}{2}}\cancel{\sigma}\, dz$$

$$= \frac{\mu}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{\frac{-z^2}{2}} dz + \frac{\sigma}{\sqrt{2\pi}}\int_{-\infty}^{\infty} 2e^{\frac{-z^2}{2}} dz$$

$\therefore \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{\frac{-z^2}{2}} dz$ represents the total probability of normal curve from with mean zero and variance 1 and hence equal to 1. The second term is zero since it represents an odd function.

$$\therefore E(X) = \mu$$

Variance of Normal Distribution

If X is a normal variate then variance of normal distribution is $E\left[(X-\mu)^2\right] = \frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\infty} (x-\mu)^2 e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx$

Let $\frac{x-\mu}{\sigma} = z$, $x = \mu + \sigma z$, $dx = \sigma dz$

$$E\left[(X-\mu)^2\right] = \frac{1}{\cancel{\sigma}\sqrt{2\pi}}\int_{-\infty}^{\infty} (\sigma z)^2 e^{\frac{-z^2}{2}}\cancel{\sigma}\, dz$$

$$= \frac{\sigma^2}{\sqrt{2\pi}}\int_{-\infty}^{\infty} z^2 e^{\frac{-z^2}{2}} dz$$

$$= \frac{\sigma^2}{\sqrt{2\pi}} ze^{\frac{-z^2}{2}}\Big|_{-\infty}^{\infty} + \int_{-\infty}^{\infty} e^{\frac{-z^2}{2}} dz$$

$$= \sigma^2(0+1) = \sigma^2$$

$\frac{z^2}{2} = t$

$\frac{\cancel{2}z}{\cancel{2}} dz = dt$

e^{-t}

Caution:

- The mean and variance of Normal distribution are called the parameters of normal distribution. X follows Normal distribution with mean μ and variance α^2 is represented as $X \sim N(\mu, \sigma^2)$.
- If $z = \frac{X-\mu}{\sigma}$ then

$$E(Z) = E\left(\frac{X-\mu}{\sigma}\right) = \frac{1}{\sigma}[E(X) - \mu]$$

$$= \frac{1}{\sigma}[\mu - \mu] = 0$$

$$V(z) = V\left(\frac{X-\mu}{\sigma}\right) = \frac{1}{\sigma^2}V(X)$$

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

$\therefore X$ is a normal variate with mean μ and variance σ^2. Hence, z is called a standard normal variate with mean zero and variance 1.

Median of Normal Distribution

If M is the mean of normal distribution, it should divide the entire distribution into two equal parts.

$$\int_{-\infty}^{M} f(x)dx = \frac{1}{2}$$

$$\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{M} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx = \frac{1}{2}$$

$$\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\mu} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx + \frac{1}{\sigma\sqrt{2\pi}}\int_{\mu}^{M} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx = \frac{1}{2}$$

However,

$$\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\mu} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{0} e^{\frac{-z^2}{2}} dz$$

$$= \frac{1}{2}\begin{cases}\because \dfrac{X-\mu}{\sigma} = z \\ dX = \sigma dZ\end{cases}$$

$$\frac{1}{2} + \frac{1}{\sigma\sqrt{2\pi}}\int_{\mu}^{M} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx = \frac{1}{2}$$

$$\frac{1}{\sigma\sqrt{2\pi}}\int_{\mu}^{M} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx = 0$$

$$\therefore \mu = \text{M}$$

Hence, for a normal distribution Median = μ

Mode of Normal Distribution

Mode is the value of x that occurs maximum number of times, that is, a value with highest frequency in a distribution. Mode is the value of x, for which $f(x)$ is maximum. Hence, mode is the solution of $f'(x) = 0$ and $f''(x) < 0$.

The normal distribution with mean μ and Standard deviation σ is

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

Taking log on both sides

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

Differentiating both sides w.r.t. 'x'

$$\frac{1}{f(x)} f'(x) = 0 - \frac{2}{2}\left(\frac{x-\mu}{\sigma}\right)\frac{1}{\sigma}$$

$$f'(x) = -f(x)\left(\frac{x-\mu}{\sigma^2}\right) \tag{5.1}$$

$$f'(x) = 0 \Rightarrow \frac{x-\mu}{\sigma^2} = 0 \qquad \{\because f(x) \neq 0\}$$

$\therefore x = \mu$ differentiating equation (5.1) w.r.t. x, we get

$$f''(x) = -\left[f'(x)\left(\frac{x-\mu}{\sigma^2}\right) + f(x)\cdot\frac{1}{\sigma^2}\right]$$

$$= -f(x)\left(\frac{x-\mu}{\sigma^2}\right)^2 + f(x)\cdot\frac{1}{\sigma^2} \qquad \{\text{Substituting for } f'(x) \text{ from equation (5.1)}\}$$

At $x = \mu$, $f''(x) = -f(x)\left[\left(\frac{x-\mu}{\sigma^2}\right)^2 + \frac{1}{\sigma^2}\right]$

$$= \frac{-1}{\sigma\sqrt{2\pi}}\cdot\frac{1}{\sigma^2} < 0 \left\{\because f(x)|_{x=\mu} = \frac{1}{\sigma\sqrt{2\pi}}\right\}$$

$\therefore f''(x) < 0$, hence $f(x)$ attains maximum at $x = \mu$, and the maximum value is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}$$

Hence mode of normal distribution is $x = \mu$.

Points of Inflection of Normal Distribution

These are the points where the curve closes the tangent.

At the points of inflection of the normal curve we have,

$$f''(x) = 0 \text{ and } f''(x) \neq 0.$$

From the previous derivations,

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

$$f'(x) = -f(x)\left(\frac{x-\mu}{\sigma^2}\right)$$

$$f''(x) = -\left[f'(x)\left[\left(\frac{x-\mu}{\sigma^2}\right)\right] + f(x)\frac{1}{\sigma^2}\right]$$

$$= +f(x)\left(\frac{x-\mu}{\sigma^2}\right)^2 - f(x)\cdot\frac{1}{\sigma^2} \tag{5.2}$$

$$= f(x)\left[\left(\frac{x-\mu}{\sigma^2}\right)^2 - \frac{1}{\sigma^2}\right]$$

$$= 0$$

$$f'''(x) = 0 \Rightarrow \left(\frac{x-\mu}{\sigma^2}\right)^2 - \frac{1}{\sigma^2} = 0 \quad \because f(x) \neq 0$$

$$\frac{(x-\mu)^2}{\sigma^4} = \frac{1}{\sigma^2}, \quad (x-\mu)^2 = \sigma^2$$

$$x - \mu = \pm\sigma$$

$$x = \mu \pm \sigma$$

Differentiating equation (5.2) w.r.t. x, we get

$$f'''(x) = f'(x)\left[\left(\frac{x-\mu}{\sigma^2}\right)^2 - \frac{1}{\sigma^2}\right] + f(x)\left[2\left(\frac{x-\mu}{\sigma^2}\right)\frac{1}{\sigma^2}\right]$$

$$f'''(x)\Big|_{x=\mu\pm\sigma} = f'(x)\Big|_{x=\mu\pm\sigma}\left\{\left(\frac{x-\mu}{\sigma^2}\right)^2 - \frac{1}{\sigma^2}\right\} + f(x)\Big|_{x=\mu\pm\sigma}\left\{\frac{2(x-\mu)}{\sigma^2}\frac{1}{\sigma^2}\right\}_{x=\mu\pm\sigma}$$

$$= f'(x)\left[\frac{1}{\sigma^2} - \frac{1}{\sigma^2}\right] + f(x)2\frac{1}{\sigma^3}$$

$$f(x)\Big|_{x=\mu\pm\sigma} = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}}$$

Hence, $$f'''(x) = 2\frac{1}{\sigma^3}\cdot\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}} \neq 0$$

Hence the points of inflection of normal distribution are $x = \mu \pm \sigma$. Since $f''(x) = 0$ and $f'''(x) \neq 0$ for that x, at $x = \mu \pm \sigma$.

Mean Deviation About Mean of Normal Distribution

$$\text{M.D. about mean} = \int_{-\infty}^{\infty} |x-\mu| f(x)\,dx$$

$$= \int_{-\infty}^{\infty} |x-\mu| \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx$$

Let $$\frac{x-\mu}{\sigma} = z$$

$$x - \mu = \sigma z$$

$$dx = \sigma\,dz$$

$$\therefore \text{M.D. about mean} = \int_{-\infty}^{\infty} |\sigma z| \cdot \frac{1}{\cancel{\sigma}\sqrt{2\pi}} e^{\frac{-z^2}{2}} dz \,(\cancel{\sigma})$$

$$= \frac{\sigma}{\sqrt{2\pi}} \int_{-\infty}^{\infty} |z| \, e^{\frac{-z^2}{2}} dz$$

The above integrand $|z| e^{\frac{-z^2}{2}}$ is an even function of z. Hence from the properties of definite integrals, we get

$$\text{M.D. about mean} = \frac{2\sigma}{\sqrt{2\pi}} \int_0^{\infty} z e^{\frac{-z^2}{2}} dz \,\{\because \text{in } (0,\infty)\ |z| = z\}$$

$$= \frac{2\sigma}{\sqrt{2\pi}} \int_0^{\infty} e^{-y} dy$$

$$= \frac{2\sigma}{\sqrt{2\pi}} \frac{e^{-y}}{-1}\Big|_0^{\infty} = \sigma\sqrt{\frac{2}{\pi}}$$

$$\cong \frac{4}{5}\sigma$$

Let $\frac{z^2}{2} = y$

$\frac{\cancel{2}z}{\cancel{2}} dz = dy$

Caution:

- $\int_{-a}^{a} f(x)dx = 0$, f is odd
- $\int_{-a}^{a} f(x)dx = 2\int_{0}^{a} f(x)dx$, f is even

Area Property of Normal Distribution

Let X be a normal variate with mean μ and variance σ^2 (i.e.,) $x \sim N(\mu,\sigma^2)$. Then the probability that the value of X will lie between $X = \mu$ and $X = x_1$ is given by

$$P(\mu < X < x_1) = \int_{\mu}^{x_1} f(x)dx$$

$$= \frac{1}{\sigma\sqrt{2\pi}} \int_{\mu}^{x_1} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx$$

Let $\frac{x-\mu}{\sigma} = z$

$x = \mu + \sigma z$

$dx = \sigma dz$

when $x = \mu$, z = 0

when $x = x_1$, $z = \frac{x_1 - \mu}{\sigma}$

$= z_1$ (say)

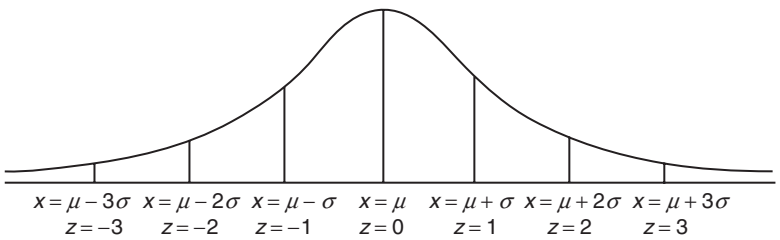
$$\therefore P(\mu < X < x_1) = P(0 < z < z_1)$$

$$= \frac{1}{\cancel{\sigma}\sqrt{2\pi}} \int_0^{z_1} e^{-z\frac{-z^2}{2}} \cancel{\sigma}\, dz$$

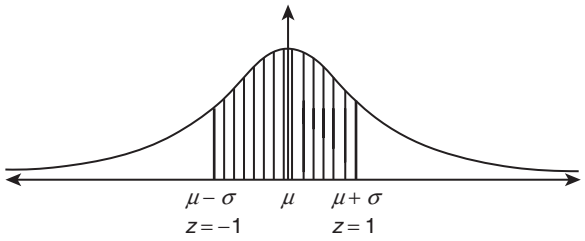
$$= \frac{1}{\sqrt{2\pi}} \int_0^{z_1} e^{\frac{-z^2}{2}} dz = \int_0^{z_1} \phi(z) dz$$

where $\phi(z) = \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}}$ is called the probability function of the standard normal variate.

The area under this normal curve is as follows:



In particular, the probability that a random value of X lies in the interval $(\mu - \sigma, \mu + \sigma)$ is given $P(\mu - \sigma < X < \mu + \sigma)$



$$\therefore P(\mu - \sigma < X < \mu + \sigma) = \int_{\mu-\sigma}^{\mu+\sigma} f(x)\, dx \quad \text{by taking } z = \frac{X - \mu}{\sigma}$$

$$\therefore P(-1 < z < 1) = \int_{-1}^{1} \phi(z)\, dz$$

$$= 2\int_0^1 \phi(z)\, dz \quad \left\{\begin{array}{l}\because \text{ normal curve is symmetric} \\ \text{about the line } x = \mu \text{ or } z = 0\end{array}\right\}$$

$$\therefore P(-1 < z < 1) = 0.6826 \quad = 2 \times 0.3413 = 0.6826$$

These areas under standard normal curve are tabulated and are given in Appendix A.

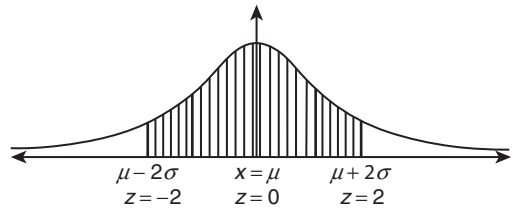
Similarly, the probability that a random value of X lies in the interval $(\mu - 2\sigma, \mu + 2\sigma)$ is

$$P(\mu - 2\sigma < x < \mu + 2\sigma) = \int_{\mu-2\sigma}^{\mu+2\sigma} f(x)\, dx$$

$$\therefore P(-2 < z < 2) = \int_{-2}^{2} \phi(z)\, dz$$

$$P(-2 < z < 2) = 2\int_0^2 \phi(z)\,dz$$
$$= 2 \times 0.4772$$
$$= 0.9544$$

$$\therefore P(-2 < z < 2) = 0.9544$$

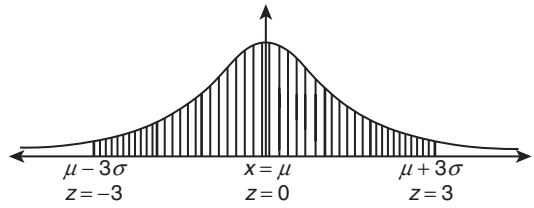


Similarly,

$$P(\mu - 3\sigma < x < \mu + 3\sigma) = \int_{\mu-3\sigma}^{\mu+3\sigma} f(x)dx$$

$$P(-3 < z < 3) = \int_{-3}^{3} \phi(z)dz = 2\int_0^3 \phi(z)\,dz$$
$$= 2 \times (0.4986) = 0.9973$$

$$\therefore P(-3 < z < 3) = 0.9973$$



The MGF of Normal Distribution

Let X be a normal variate. Then MGF (about origin) is given by

$$M_X(t) = E(e^{tx}) = \int_{-\infty}^{\infty} e^{tx} f(x)\,dx$$
$$= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{tx} e^{\frac{-1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx$$

$$\text{Let } z = \frac{X-\mu}{\sigma} \quad x = \mu + \sigma z$$
$$dx = \sigma dz$$

$$M_X(t) = \frac{1}{\cancel{\sigma}\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{t(\mu+\sigma z)} e^{-\frac{1}{2}z^2} \cancel{\sigma}\,dz$$
$$= \frac{e^{\mu t}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(z^2+\sigma^2t^2-2t\sigma z)} e^{\frac{\sigma^2t^2}{2}} dz$$

$$\text{Let } \frac{z-\sigma t}{\sqrt{2}} = y$$
$$dz = (dy)\sqrt{2}$$

$$= \frac{e^{\mu t}}{\sqrt{2\pi}} e^{\frac{\sigma^2t^2}{2}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(z-\sigma t)^2} dz$$

$$= \frac{e^{\mu t + \frac{\sigma^2t^2}{2}}}{\cancel{\sqrt{2\pi}}} \int_{-\infty}^{\infty} e^{-y^2} dy \cancel{\sqrt{2}} = \frac{e^{\mu t + \frac{\sigma^2t^2}{2}}}{\sqrt{\pi}} 2\int_0^{\infty} e^{-y^2} dy$$

$$= e^{\mu t + \frac{\sigma^2t^2}{2}} \cdot \frac{\sqrt{\pi}}{\sqrt{\pi}} \quad \left\{\because 2\int_0^{\infty} e^{-y^2} dy = \sqrt{\pi}\right\}$$

$$= e^{\mu t + \frac{\sigma^2t^2}{2}}$$

Caution:

- The MGF of standard normal variate,

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

$$M_z(t) = M_{\frac{X-\mu}{\sigma}}(t) = e^{\frac{-\mu}{\sigma}} M_X\left(\frac{t}{\sigma}\right)$$

$$E(tz) = E\left[e^{t\frac{(X-\mu)}{\sigma}}\right] = E\left[e^{\frac{-t\mu}{\sigma}} e^{\frac{tx}{\sigma}}\right] = e^{\frac{-t\mu}{\sigma}} M_X\left(\frac{t}{\sigma}\right)$$

$$= e^{\frac{-t\mu}{\sigma}} e^{\mu t} = e^{\frac{t^2}{2}}$$

Moments of Normal Distribution

We shall derive a separate formula for odd and even ordered moments.
Odd ordered moments about mean are given by

$$\mu_{2r+1} = \int_{-\infty}^{\infty} (x-\mu)^{2r+1} f(x)dx$$

$$= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} (x-\mu)^{2r+1} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx$$

$$\text{Let } z = \frac{X-\mu}{\sigma}$$

$$X - \mu = \sigma z$$

$$dX = \sigma\, dz$$

$$\therefore \mu_{2r+1} = \frac{1}{\cancel{\sigma}\sqrt{2\pi}} \int_{-\infty}^{\infty} (\sigma z)^{2r+1} e^{\frac{-z^2}{2}} \cancel{\sigma}\, dz$$

$$= \frac{\sigma^{2r+1}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z^{2r+1} e^{\frac{-z^2}{2}} dz$$

The integrand above is an odd function of z and hence from the property of definite integrals, we get

$$\mu_{2r+1} = 0 \tag{5.3}$$

Even ordered moments about mean are given by

$$\mu_{2r} = \int_{-\infty}^{\infty} (x-\mu)^{2r} f(x)dx$$

$$= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} (x-\mu)^{2r} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx$$

$$\therefore \mu_{2r} = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} (\sigma z)^{2r} e^{\frac{-z^2}{2}} \qquad \text{Let } z = \frac{X-\mu}{\sigma}$$

$$X - \mu = \sigma z$$

$$= \frac{\sigma^{2r}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z^{2r} e^{\frac{-z^2}{2}} dz \qquad dx = \sigma\, dz$$

$$\mu_{2r} = \frac{\sigma^{2r}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z^{2r} e^{\frac{-z^2}{2}} dz$$

$$= \frac{2\sigma^{2r}}{\sqrt{2\pi}} \int_{0}^{\infty} (z^2)^r e^{\frac{-z^2}{2}} dz \quad \left\{\begin{array}{l}\text{Since the integrand is an}\\ \text{even function of } z\end{array}\right\} \qquad \text{Let } t = \frac{z^2}{2}$$

$$dt = \frac{2z}{2} dz$$

$$= \frac{2\sigma^{2r}}{\sqrt{2\pi}} \int_{0}^{\infty} (2t)^r e^{-t} \frac{dt}{\sqrt{2t}}$$

$$\mu_{2r} = \frac{\sigma^{2r} 2^r}{\sqrt{\pi}} \int_{0}^{\infty} t^{r-\frac{1}{2}} e^{-t} dt \tag{5.4}$$

Change r to $(r-1)$

$$\mu_{2r-2} = \frac{\sigma^{2r-2} 2^{r-1}}{\sqrt{\pi}} \int_{0}^{\infty} t^{r-\frac{1}{2}-1} e^{-t} dt \tag{5.5}$$

We know from the definition of Gamma function

$$\Gamma(n) = \int_{0}^{\infty} e^{-x} x^{n-1} dx$$

Also, $\Gamma(n) = (n-1)\,\Gamma(n-1)$

$$\therefore \mu_{2r-2} = \frac{\sigma^{2r-2} 2^{r-1}}{\sqrt{\pi}} \Gamma\left(r - \frac{1}{2}\right) \text{ from equation (5.5)}$$

Equation (5.4) gives,

$$\mu_{2r} = \frac{\sigma^{2r} 2^r}{\sqrt{\pi}} \Gamma\left(r - \frac{1}{2} + 1\right)$$

$$= \frac{\sigma^{2r} 2^r}{\sqrt{\pi}} \Gamma\left(r + \frac{1}{2}\right)$$

Dividing the above two equations one by the other, we get

$$\frac{\mu_{2r}}{\mu_{2r-2}} = 2\sigma^2 \frac{\Gamma\left(r + \frac{1}{2}\right)}{\Gamma\left(r - \frac{1}{2}\right)} = 2\sigma^2 \left(r - \frac{1}{2}\right) \text{ from the property of } \Gamma(n)$$

$$\mu_{2r} = \sigma^2 (2r-1)\, \mu_{2r-2} \tag{5.6}$$

This is recurrence relation for the moments of normal distribution.
Using the recurrence relation for μ_{2r-2} we get,

$$\mu_{2r} = [\sigma^2(2r-1)][(2r-3)\sigma^2]\mu_{2r-4}$$

$$= [\sigma^2(2r-1)][(2r-3)\sigma^2][(2r-5)\sigma^2]\mu_{2r-6}$$

Proceeding in this way, we get

$$\begin{aligned}\mu_{2r} &= \sigma^{2r}(2r-1)(2r-3)(2r-5)\ldots 5\cdot 3\cdot 1\mu_0 \\ &= \sigma^{2r}[1\cdot 3\cdot 5\ldots(2r-5)(2r-3)(2r-1)]\end{aligned} \tag{5.7}$$

Hence, odd ordered moments of normal distribution vanish whereas even ordered moments are given by equation (5.7).

Alternative Method

These moments can also be generated from the MGF about mean as follows:

The MGF about mean is given by

$$\begin{aligned}&= E[e^{t(X-\mu)}] = E[e^{tX}e^{-\mu t}] = e^{-\mu t}E[e^{tX}] \\ &= e^{-\mu t}M_X(t) \\ &= e^{-\mu t}[\text{MGF about origin}] \\ &= e^{-\mu t}\left[e^{\mu t} + t^2\frac{\sigma^2}{2}\right] \\ &= et^2\frac{\sigma^2}{2} \\ &= 1+\left(t^2\frac{\sigma^2}{2}\right)+\frac{\left(t^2\frac{\sigma^2}{2}\right)^2}{2!}+\frac{\left(t^2\frac{\sigma^2}{2}\right)^3}{3!}+\ldots+\frac{\left(t^2\frac{\sigma^2}{2}\right)^n}{n!}+\ldots\end{aligned} \tag{5.8}$$

The coefficient of $\frac{t^r}{r!}$ in the above equation gives r^{th} moment about mean, namely μ_r. We can observe from equation (5.8) that there are no terms with t^r. Hence coefficient of $\frac{t^r}{r!} = 0$

$$\therefore \mu_{2n+1} = 0, \quad n = 0, 1, 2, \ldots$$

The coefficient of $\frac{t^{2r}}{(2r)!}$ in equation (5.8) gives

$$\begin{aligned}\mu_{2n} &= \text{coefficient of } \frac{t^{2r}}{(2r)!} \text{ in (5.9)} \\ &= \frac{\sigma^{2n}(2n)!}{2^n n!} \\ &= \frac{\sigma^{2n}}{2^n n!}[(2n)(2n-1)(2n-2)\ldots 5\cdot 4\cdot 3\cdot 2\cdot 1] \\ &= \frac{\sigma^{2n}}{2^n n!}[(2n)(2n-2)(2n-4)\ldots 4\cdot 2][(2n-1)(2n-3)\ldots 5\cdot 3\cdot 1] \\ &= \frac{\sigma^{2n}}{2^n n!}[2^n n(n-1)(n-2)\ldots 2\cdot 1][1\cdot 3\cdot 5\cdots(2n-3)(2n-1)] \\ &= \frac{\sigma^{2n}}{2^n n!}2^n n![1\cdot 3\cdot 5\cdots(2n-3)(2n-1)] \\ &= \sigma^{2n}[1\cdot 3\cdot 5\cdots(2n-1)]\end{aligned}$$

In the above topics, we have derived some of the chief characteristics of normal distribution. Many a times, we come across a situation where we need a normal approximation to a binomial distribution. In the next topic, let us discuss the normal approximation to a binomial distribution.

Normal Distribution as a Limiting Form of Binomial Distribution

The following assumptions are required to obtain the limiting form of binomial distribution to normal distribution:

(i) n, the number of trials is indefinitely large, (i.e.,) $n \to \infty$

(ii) Neither p nor q is very small.

Consider the probability mass function of the binomial distribution with parameters n and p.

$$\begin{aligned} p(x) &= n_{C_x} p^x q^{n-x}, \qquad x = 0,1,2,\ldots n \\ &= \frac{n!}{x!(n-x)!} p^x q^{n-x} \end{aligned} \tag{5.9}$$

To find the limiting form of the above equation (5.9), we use Stirling's approximation to $n!$ for large n.

$$\underset{n\to\infty}{Lt}\ n! \cong \sqrt{2\pi} e^{-n} r^{n+\left(\frac{1}{2}\right)}$$

$x \to \infty$ follows consequently as $n \to \infty$.

Substituting this approximation in equation (5.9), we get

$$\begin{aligned} \lim\ p(x) &= \frac{\cancel{\sqrt{2\pi}}\ \cancel{e^{-n}}\ n^{n+\left(\frac{1}{2}\right)} p^x q^{n-x}}{\cancel{\sqrt{2\pi}}\ \cancel{e^{-x}}\ x^{x+\left(\frac{1}{2}\right)} \cancel{\sqrt{2\pi}}\ \cancel{e^{-(n-x)}} (n-x)^{n-x+\frac{1}{2}}} \\ &= \frac{n^{x+\frac{1}{2}} p^{x+\frac{1}{2}} n^{n-x+\frac{1}{2}} q^{n-x+\frac{1}{2}}}{\sqrt{2\pi} x^{x+\frac{1}{2}} (n-x)^{n-x+\frac{1}{2}} n^{\frac{1}{2}} p^{\frac{1}{2}} q^{\frac{1}{2}}} \\ &= \frac{1}{\sqrt{2\pi}} \cdot \frac{1}{\sqrt{npq}} \frac{(np)^{x+\frac{1}{2}} (nq)^{n-x+\frac{1}{2}}}{x^{x+\frac{1}{2}} (n-x)^{n-x+\frac{1}{2}}} \\ &= \frac{1}{\sqrt{2\pi}} \cdot \frac{1}{\sqrt{npq}} \left(\frac{np}{x}\right)^{x+\frac{1}{2}} \left(\frac{nq}{n-x}\right)^{n-x+\frac{1}{2}} \end{aligned} \tag{5.10}$$

Consider the standard normal variate if X is a normal variate.

$$z = \frac{X-\mu}{\sigma} = \frac{X - E(X)}{\sqrt{V(X)}}$$

Consider the Standard binomial variate, where $E(X) = np$ and $V(X) = npq$ if X is a binomial variate.

$$Z = \frac{X-np}{\sqrt{npq}}, \qquad X = 0, 1, 2, \ldots n \tag{5.11}$$

When $X = 0$, $z = \dfrac{-np}{\sqrt{npq}} = -\sqrt{\dfrac{np}{q}}$

When $X = n$, $z = \dfrac{n-np}{\sqrt{npq}} = \dfrac{nq}{\sqrt{npq}} = \sqrt{\dfrac{nq}{p}}$

In the limit form as $n \to \infty$, z takes the values from $-\infty$ to ∞.
From equation (5.11)

$$z\sqrt{npq} = X - np$$
$$X = np + z\sqrt{npq} \tag{5.12}$$

$$\frac{X}{np} = 1 + z\sqrt{\frac{q}{np}} \tag{5.13}$$

$$n - X = n - np - z\sqrt{npq} = nq - z\sqrt{npq}$$
$$\frac{n-X}{nq} = 1 - z\sqrt{\frac{p}{nq}} \tag{5.14}$$

From equation (5.11) $dz = \frac{1}{\sqrt{npq}}\,dx$

The probability differential of the distribution of z in the limit is

$$dG(z) = g(z)dz = \underset{n\to\infty}{Lt}\left[\frac{1}{\sqrt{2\pi}} \times \frac{1}{N}\right]dz \qquad \text{(5.15) from equation (5.2)}$$

where

$$N = \left(\frac{x}{np}\right)^{x+\frac{1}{2}}\left(\frac{n-x}{nq}\right)^{n-x+\frac{1}{2}}$$

Applying logarithm on both sides we get

$$\log N = \left(x + \frac{1}{2}\right)\log\left(\frac{x}{np}\right) + \left(n - x + \frac{1}{2}\right)\log\left(\frac{n-x}{nq}\right)$$

Substituting for x in terms of z we get from equation (5.12)

$$\log N = \left[np + z\sqrt{npq} + \frac{1}{2}\right]\log\left[1 + z\sqrt{\frac{q}{np}}\right]$$
$$+ \left[n - np - z\sqrt{npq} + \frac{1}{2}\right]\log\left[1 - z\sqrt{\frac{p}{nq}}\right]$$

$$= \left[np + z\sqrt{npq} + \frac{1}{2}\right]\left[z\sqrt{\frac{q}{np}} - \frac{1}{2}z^2\left(\frac{q}{np}\right) + \frac{1}{3}z^3\left(\frac{q}{np}\right)^{\frac{3}{2}}\cdots\right]$$
$$+ \left[np - z\sqrt{npq} + \frac{1}{2}\right]\left[-z\sqrt{\frac{p}{nq}} - \frac{1}{2}z^2\left(\frac{p}{nq}\right) - \frac{1}{3}z^3\left(\frac{p}{nq}\right)^{\frac{3}{2}}\cdots\right]$$

$$= \left[\cancel{z\sqrt{npq}} - \frac{1}{2}z^2q + z^2q - \frac{1}{2}z^3q^{\frac{3}{2}} + \frac{1}{2}z\sqrt{\frac{q}{np}} - \frac{1}{4}z^2\left(\frac{q}{np}\right)\right]$$
$$\left[-\cancel{z\sqrt{npq}} - \frac{1}{2}z^2p + z^2p - \frac{1}{2}z^3p^{\frac{3}{2}} - \frac{1}{2}z\sqrt{\frac{p}{nq}} - \frac{1}{4}z^2\left(\frac{p}{nq}\right)\cdots\right]$$

$$= \frac{-1}{2}z^2(p+q) + z^2(p+q) - \frac{1}{2}z^3\left(p^{\frac{3}{2}} + q^{\frac{3}{2}}\right) + \frac{z}{2\sqrt{n}}\left(\sqrt{\frac{q}{p}} + \sqrt{\frac{p}{q}}\right) + 0\left(n^{\frac{-1}{2}}\right)$$

$$= \frac{-z^2}{2} + z^2 + 0\left(n^{\frac{-1}{2}}\right)$$

$$= \frac{z^2}{2} \text{ as } n \to \infty$$

$$\underset{n\to 0}{Lt} \log N = \frac{z^2}{2} \qquad \therefore \underset{n\to\infty}{Lt} N = e^{\frac{z^2}{2}}$$

Substituting this in equation (5.15) we get

$$dG(z) = g(z)dz = \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}} dz$$

Hence the probability function of z is

$$g(z) = \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}}, \qquad -\infty < z < \infty$$

The above function is called pdf of the standard normal distribution with mean 0 and variance 1.

Importance and Applications of Normal Distribution

The major applications of normal distribution are statistical quality control, testing of significance, sampling distributions of various statistics, graduation of the non-normal curves, etc. Some more examples where normal distribution can be covered are weights taken for a group of students (taken of the same age, intelligence), proportion of male to female births for some particular geographical region over a period of years, etc.

Normal distribution is very important in statistical theory because of the following reasons:

(i) The distributions like Binomial, Poisson, Hypergeometric distributions, etc., can be approximated by normal distributions.

(ii) Many of the sampling distributions like student's t, Snedecor's F, Chi-square distributions etc., tend to normality for large samples.

(iii) The entire theory of small sample tests like t, F, χ^2 tests etc., is based on the fundamental assumption that the parent populations from which the samples have been drawn follow Normal distribution.

(iv) Even if a variable is not normally distributed, it can sometimes be brought to normal form by transformation of variable.

(v) The area property of normal distribution forms the basis of entire large sample theory.

(i.e.,) If $x \sim N(\mu, \sigma^2)$ then

$$P(\mu - 3\sigma \le X \le \mu + 3\sigma) = P(-3 < z < 3) = 0.9973$$

$$\therefore P(|z| > 3) = 1 - P(|z| \le 3)$$

$$= 1 - 0.9973$$

$$= 0.0027$$

(vi) W. J. Youden of the National Bureau of Standards describes the importance of normality distribution as 'The normal law of errors stands out in the experience of mankind as one of broadest generalizations of natural philosophy. It serves as the guiding instrument in researches, in the physical and social sciences and in medicine, agriculture and engineering. It is indispensible tool for the analysis and the interpretation of the basic data obtained by observation and experiment'.

Fitting of Normal Distribution

The first step in fitting of the distribution to the given data is to calculate the mean μ and standard deviation σ from the given data. Then the normal curve that is fit to the given data is given by

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

The expected frequencies are calculated using the formula $E(X) = NP(x)$, where $P(X_1 < X < X_2) = P(z_1 < z < z_2)$.

Using $z_1 = \frac{X_1 - \mu}{\sigma}, z_2 = \frac{X_2 - \mu}{\sigma}$ and the areas are obtained using normal tables. This is illustrated with the examples given in the next section.

Worked Out Examples

EXAMPLE 5.1

Fit a normal distribution to the following data:

Class:	60–62	63–65	66–68	69–71	72–74
Frequency:	5	18	42	27	8

Solution: The total frequency $N = 100$
First the mean μ and Standard deviation σ are calculated as follows:

C.I	f_i	Mid value (x_i)	$f_i x_i$	$(x_i - \mu)^2$	$f_i(x_i - \mu)^2$
60–62	5	61	305	41.6025	208.0125
63–45	18	64	1152	11.9025	214.245
66–68	42	67	2814	0.2025	8.505
69–71	27	70	1890	6.5025	175.5675
72–74	8	71	584	30.8025	246.42
	$N = 100$		$N = 6745$		852.75

$$\text{Mean } \mu = \frac{\Sigma f_i x_i}{\Sigma f_i}$$

$$= \frac{6745}{100} = 67.45$$

$$\text{Standard deviation } \sigma = \sqrt{\frac{\Sigma f_i (x_i - \mu)^2}{N}}$$

$$= \sqrt{\frac{852.45}{100}}$$

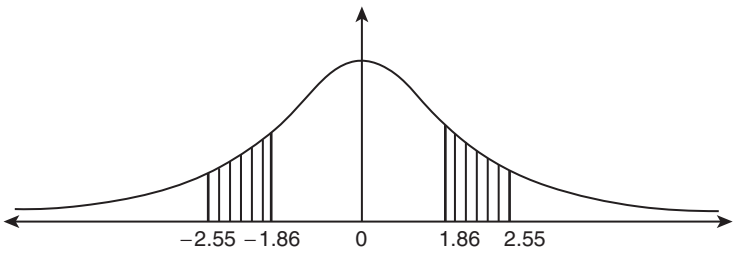
$$= 2.92$$

Calculation of Expected Frequencies

$$P(60 < X < 62) = P(-2.55 < z < -1.86)$$

$$X = 60,\ z = \frac{60 - 67.45}{2.92} = -2.55$$

$$X = 62,\ z = \frac{62 - 67.45}{2.92} = -1.86$$



$$P(-2.55 < z < -1.86) = P(1.86 < z < 2.55)$$

$$= P(0 < z < 2.55) - P(0 < z < 1.86)$$

$$= 0.9946 - 0.9686 = 0.026$$

$$\therefore E(X_1) = NP\ (60 < X < 62) = 100\ (0.026) = 2.6$$

Class	Lower class $z = \left(\frac{X - \mu}{\sigma}\right)$	Upper class	$P(z_1 < z < z_2)$	$E(x) = NP(z_1 < z < z_2)$
60–62	$z_1 = -2.55$	$z_2 = -1.86$	0.4946 − 0.4686 = 0.026	100(0.026) = 2.6
63–65	$z_1 = -1.52$	$z_2 = -0.84$	0.4357 − 0.4995 = 0.1362	100(0.1362) = 13.62
66–68	$z_1 = -0.49$	$z_2 = 0.18$	0.0714 + 0.1879 = 0.2593	100(0.2593) = 25.93
69–71	$z_1 = 0.53$	$z_2 = 1.215$	0.3808 − 0.2019 = 0.1859	100(0.185) = 18.59
72–74	$z_1 = 1.55$	$z_2 = 2.24$	0.4875 − 0.4394 = 0.0481	100(0.0481) = 4.81

The expected frequencies are given by

C.I	60–62	63–65	66–68	69–71	72–74
Ex. Frequency	3	14	26	19	5

Worked Out Problems

EXAMPLE 5.2

Find the area under the normal curve in each of following:

(i) $z = 0$ and $z = 1.2$

(ii) $z = -0.68$ and $z = 0$

(iii) $z = -0.46$ and $z = 2.21$

(iv) $z = 0.81$ and $z = 1.94$

(v) to the left of $z = -0.6$

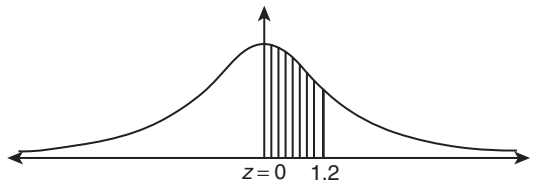
(vi) to right of $z = 1.28$.

Solution:

(i) $P(z = 0 \text{ and } z = 1.2) = P(0 < z < 1.2)$

$= P(0 < z < 1.2)$

$= 0.3849$

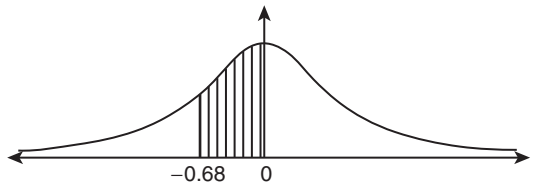


(ii) $P(z = -0.68 \text{ and } z = 0) = P(-0.68 < z < 0)$

By symmetry of normal curve we have

The above area $= P(0 < z < 0.68)$

$= 0.2517$



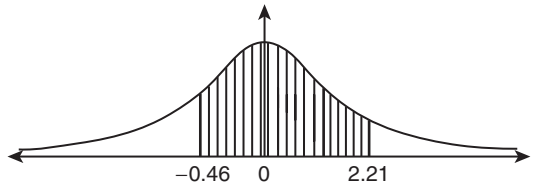
(iii) $P(z = -0.46 \text{ and } z = 2.21)$

$= P(0 < z < 0.46) + P(0 < z < 2.21)$

{Since by symmetry of normal curve we have $P(-0.46 < z < 0) = P(0 < z < 0.46)$}

Required area $= 0.1772 + 0.4874$

$= 0.6636$

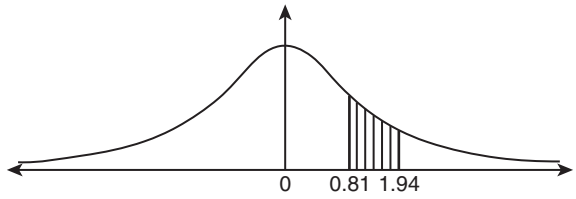


(iv) $P(z = 0.81 \text{ and } z = 1.94)$

$= P(0 < z < 1.94) - P(0 < z < 0.81)$

$= 0.4738 - 0.2910$

$= 0.1828$



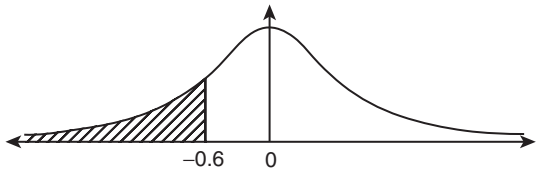
(v) P(to left of $z = -0.6$)

$= P(-\infty < z < -0.6)$

$= P(0.6 < z < \infty)$

$= P(0 < z < \infty) - P(0 < z < 0.6)$

$= 0.5 - 0.2257 = 0.2743$

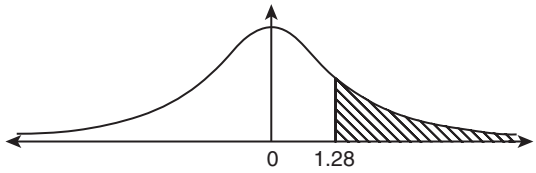


(vi) P(right of $z = 1.28$) $= P(1.28 < z < \infty)$

$= P(0 < z < \infty) - P(0 < z < 1.28)$

$= 0.5 - 0.3997$

$= 0.1003$



EXAMPLE 5.3

X is a normal variate with mean 30 and standard deviation 5. Find the probability that

(i) $26 \le X \le 40$

(ii) $X \le 45$

(iii) $|X - 30| > 5$.

Solution:

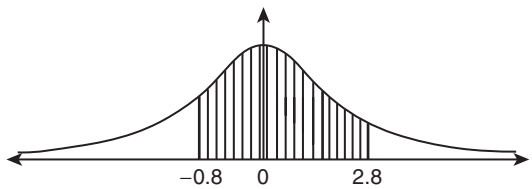
(i) $P(26 \le X \le 40)$:

When $X = 26$, $z = \dfrac{X - \mu}{\sigma} = \dfrac{26 - 30}{5}$

$= -0.8$

Given $\mu = 30$, $\sigma = 5$

When $X = 40$, $z = \dfrac{40 - 26}{5} = 2.8$



$P(26 \le X \le 40) = P(-0.8 < z < 2.8)$

$= P(0 < z < 0.8) + P(0 < z < 2.8)$

$= 0.2881 + 0.4974$

$= 0.7855$

(ii) $P(X \ge 45)$

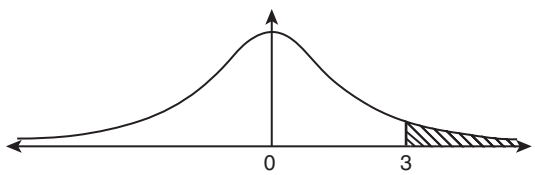
When $X = 45$, $z = \dfrac{45 - 30}{5} = 3$

$P(X \ge 45) = P\ (z \ge 3)$

$= P(0 < z < \infty) - P(0 < z < 3)$

$= 0.5 - 0.4987$

$= 0.0013$

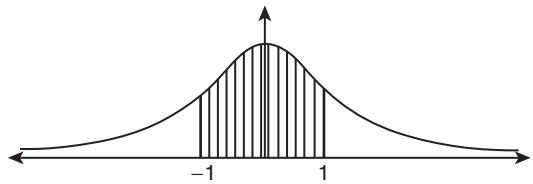


(iii) $P\left[|X-30|>5\right]=1-P\left[|X-30|\le 5\right]$

$$P\left[|X-30|\le 5\right]=P(-5\le X-30\le 5)=P(25\le X\le 35)$$

$$X=25,\ z=\frac{25-30}{5}=-1$$

$$X=35,\ z=\frac{35-30}{5}=1$$



$$P(|x-30|)\le 5=P(-1\le z\le 1)$$
$$=2P\ (0\le z\le 1)$$
$$=2(0.3413)$$
$$=0.6826$$
$$P(|X-30|)>5=1-P[|X-30|\le 5]$$
$$=1-0.6826$$
$$=0.3714$$

EXAMPLE 5.4

A random variable has a normal distribution with $\sigma = 10$. Of the probability is 0.8212 that it will take on a value less than 82.5, what is the probability that it will take on a value greater than 58.3.

Solution: Given $\sigma = 10$

$$P(X\le 82.5)=0.8212$$

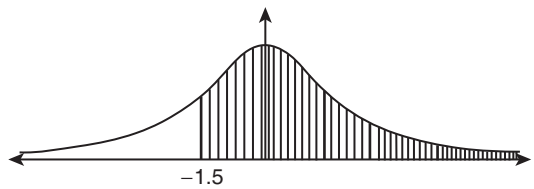
Let $$z=\frac{X-\mu}{\sigma},\ P\left(\frac{82.5-\mu}{10}\right)=0.8212,$$

From normal tables,

$$z=0.92$$

$$0.92=\frac{82.5-\mu}{10}\Rightarrow \mu=73.3$$

Probability that it will take on a value greater than $58.3 = P(X \ge 58.3)$



$$X=58.3$$

$$z=\frac{58.3-73.3}{10}=-1.5$$

$$\therefore P(X\ge 58.3)=P(z\ge -1.5)$$
$$=P(-1.5\le z<0)+P(0\le z\le \infty)$$
$$=0.4332+0.5$$
$$=0.9332$$

EXAMPLE 5.5

Suppose the heights of Indian men approximately are normally distributed with mean $\mu = 68$ and standard deviation $\sigma = 2.5$. Find the percentage of Indian men who are

(i) Between $a = 66$ and $b = 71$ inches tall

(ii) Between $a = 69.5$ and $b = 70.5$ inches tall

(iii) At least 72 inches tall.

Solution: Given that $\mu = 68$ and $\sigma = 2.5$.

Let X be height of Indian men.

Probability of heights between 66.0 and 77 inches = $P(66 < X < 71)$

When $X = 66$, $z = \dfrac{X - \mu}{\sigma}$

$= \dfrac{66 - 68}{2.5}$

$= -0.8$

When $X = 71$, $z = \dfrac{X - \mu}{\sigma} = \dfrac{71 - 68}{2.5}$

$= 1.2$

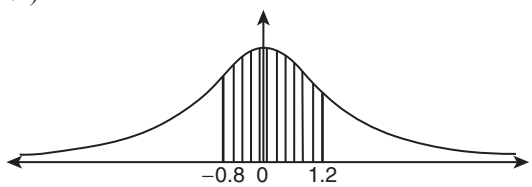
(i) $P(66 < X < 77) = P(-0.8 < z < 1.2)$

$= P(-0.8 < z < 0) + P(0 < z < 1.2)$

$= P(0 < z < 0.8) + P(0 < z < 1.2)$

$= 0.2881 + 0.3849$

$= 0.673$



∴ Percentage of men with heights between 66 and 71 inches = 100(0.673) = 67.3

(ii) Probability of height between 69.5 and 70.5 inches = $P(69.5 < X < 70.5)$

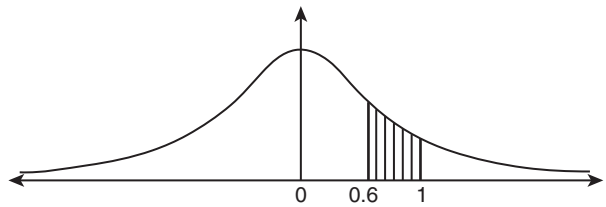
When $X = 69.5$, $z = \dfrac{X - \mu}{\sigma}$

$= \dfrac{69.5 - 65}{2.5}$

$= 0.6$

When $X = 70.5$, $z = \dfrac{70.5 - 68}{2.5}$

$= 1$



$P(69.5 < X < 70.5) = P(0.6 < z < 1)$

$= P(0 < z < 1) - P(0 < z < 0.6)$

$= 0.3413 - 0.2257$

$= 0.1156$

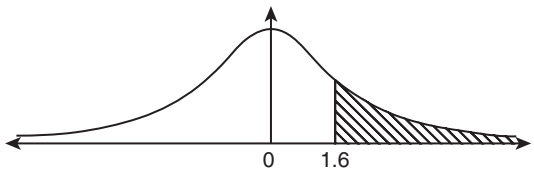
∴ Percentage of men with height lying between 69.5 and 70.5 inches is 100 (0.1156) = 11.56.

(iii) Probability of height of men at least 72 inches tall

$$= P(X \geq 72)$$

When $X = 72$, $z = \dfrac{X - \mu}{\sigma} = \dfrac{72 - 68}{2.8} = 1.6$

$$P(x \geq 72) = P(z \geq 1.6)$$
$$= P(0 < z < \infty) - P(0 < z < 1.6)$$
$$= 0.5 - 0.4452$$
$$= 0.0548$$



Percentage of men with height at least 72 inches tall

$$= 100 \times (0.0548)$$
$$= 5.48$$

EXAMPLE 5.6

The mean intelligence quotient (I.Q) of a large number of children of age 14 was 100 and the standard deviation 16. Assuming that the distribution was normal find, the following:

(i) What percentage of the children had I.Q under 80?

(ii) Between what limits, the I.Q's of the middle 40% of children lie.

Solution: Given that mean I.Q = 100 = μ

Standard deviation, $\sigma = 16$

(i) Let Xdenotes the I.Q of children of age 14. Probability that children had I.Q under 80

$$= P(X < 80)$$

When $X = 80$, $z = \dfrac{X - \mu}{\sigma} = \dfrac{80 - 100}{16} = -1.25$

$$P(X < 80) = P(z < -1.25)$$
$$= P(-\infty < z < -1.25)$$
$$= P(1.25 < z < \infty)$$
$$= P(0 < z < \infty) - P(0 < z < 1.25)$$
$$= 0.5 - 0.3944$$
$$= 0.1056$$

Percentage of children whose I.Q is under 80 is 100(0.1056) = 10.56.

(ii) Let the area from $-z_{\frac{\alpha}{2}}$ to $z_{\frac{\alpha}{2}}$ as shown be 0.4, which amounts 40% of the children lie.

From the graph it is clear that

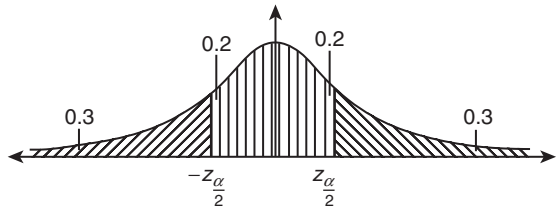
$P(0 < z < z_{\frac{\alpha}{2}}) = 0.2$

$\therefore z_{\frac{\alpha}{2}} = 0.52$

However $z = \dfrac{X - \mu}{\sigma}$

$$0.52 = \frac{X - 100}{16}$$

$$\therefore\ x = 108.32$$



Hence the I.Q of students who are in the middle 40% are $x = 108.32$.

EXAMPLE 5.7

The mean yield of one-acre plot is 662 kilos with a standard deviation of 32 kilos. Assuming normal distribution, how many one-acre plots in a batch of 1000 plots would you expect to have yield

(i) over 700 kilos

(ii) below 650 kilos and

(iii) What is the lowest yield of the best 100 plots?

Solution: Let X denote the yield in kilos for one-acre plot.

Given that, mean yield $\mu = 662$

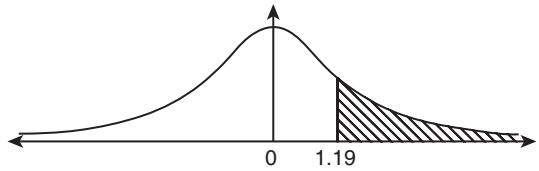
Standard deviation, $\sigma = 32$

(i) Probability that a plot will have a yield over 700 kilos = $P(X > 700)$

Let $z = \dfrac{X - \mu}{\sigma}$

When $X = 700$, $z = \dfrac{700 - 662}{32}$

$= 1.1875 \cong 1.19$



$$P(X > 700) = P(z > 1.119)$$

$$= P(1.19 < z < \infty)$$

$$= P(0 < z < \infty) - P(0 < z < 1.19)$$

$$P(X > 700) = 0.5 - 0.3830$$

$$= 0.1170$$

Hence in a batch of 1000 plots the expected number of one-acre plot with yield over 700 kilos is $1000 \times (0.1170) = 117$.

(ii) Probability that the yield of one-acre plots is below 650 kilos = $P(x < 650)$

When $X = 650$, $z = \dfrac{X - \mu}{\sigma} = \dfrac{650 - 662}{32}$

$= -0.375$

$$P(\mathrm{X} < 650) = P(Z < -0.375)$$

$$= P(-\infty < Z < -0.375)$$

$$= P(0.375 < Z < \infty)$$

$$= P(0 < Z < \infty) - (0 < Z < 0.3758)$$

$$= 0.5 - 0.14615$$

$$= 0.3538$$

Hence in a batch of 1000 plots the expected number of one-acre plots with yield below 650 kilos is $1000 \times (0.3538) = 353.8 \cong 354$.

(iii) Let the lowest yield be X_1 of best 100 plots

$$\text{(i.e.,) } P(x > x_1) = \frac{100}{1000} = 0.1$$

When $X = X_1$, $z = \dfrac{X_1 - \mu}{\sigma} = \dfrac{X_1 - 662}{32} = z_1 \text{(say)}$ (5.16)

Such that $P(z > z_1) = 0.1$

$$\Rightarrow P(0 < z < z_1) = 0.4$$

From normal tables,

$$z_1 = 1.28$$

On substituting z_1 in equation (5.16), we get

$$X_1 = 662 + 32z_1 = 662 + 32(1.28)$$

$$= 702.96$$

Hence the best 100 plot have yield over 702.96 kilos.

EXAMPLE 5.8

In an intelligence test administrated to 1,000 students the average score was 42 and the standard deviation was 24. Find

(i) The number of students exceeding a score of 50

(ii) The number of student lying between 30 and 54

(iii) The value of score exceeds by the top 100 students.

Solution: Let x denote the score in an intelligence test. Given that mean score $\mu = 42$

Standard deviation $\sigma = 24$.

(i) Probability of the score of student exceeding 50

$$= P(X > 50)$$

When $X = 50$, $z = \dfrac{x - \mu}{\sigma} = \dfrac{50 - 42}{24}$

$$= 0.33$$

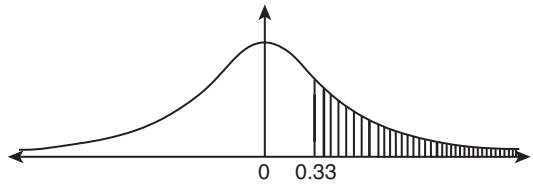
$$P(X > 50) = P(z > 0.33)$$

$= P(0.33 < z < \infty)$

$= P(0 < z < \infty) - P(0 < z < 0.31)$

$= 0.5 - 0.1293$

$= 0.3707$



Hence number of students exceeding a score of 50

$= 1000 \times 0.3707$

$= 370.7 \cong 371$

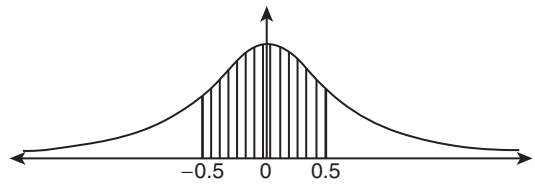
(ii) Probability that score of students lies between 30 and 54 = $P(30 < x < 54)$

When $X = 30$, $z = \dfrac{30 - 42}{24}$

$= -0.5$

When $X = 54$, $z = \dfrac{54 - 42}{24}$

$= 0.5$



$P(30 < X < 54) = P(-0.5 < z < 0.5)$

$= 2P(0 < z < 0.5)$

$= 2(0.1915)$

$= 0.383$

Hence number of students whose score lies between 30 and 54 = 1000 (0.383)

$= 383$

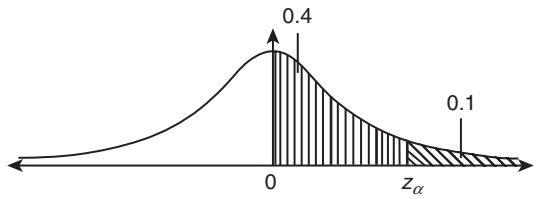
(iii) Probability of top 100 students among 1000 students $= \dfrac{100}{1000} = 0.1$

The area from z_α to ∞ (to the right of z_α) is 0.1. The statistic z_α can be found tables whose area is 0.4 (i.e.,) $P(0 < z < z_\alpha) = 0.4$

$\therefore z_\alpha = 1.28$

$$z = \frac{x - \mu}{\sigma}$$

$$1.28 = \frac{x - 42}{24} \Rightarrow x = 72.72$$



Hence the top 100 students exceed the score $x = 72.7$ (i.e.) this is the minimum score of top 100 students.

EXAMPLE 5.9

The results of a particular examination are as follows:

Students passing with distinction are 10%, students who passed are 60% and students who failed are 30%. A student is said to have failed if he gets less than 40 marks out of 100, while he has to get at

least 75 marks to pass with distinction. Determine the mean and standard deviation of the distribution of marks assuming this to be normal.

Solution: We have to find mean and standard deviation of the distribution of marks.

Let X denote the mark obtained by students in a particular examination.

Given that 30% of the students failed (i.e.,) who secured marks less than 40.

$$P(x < 40) = 30\% = 0.3$$

$$P(-\infty < z < z_\alpha) = 0.3$$

$$P(z_\alpha < z < \infty) = 0.3$$

$$P(0 < z < \infty) - P(0 < z < z_\alpha) = 0.3$$

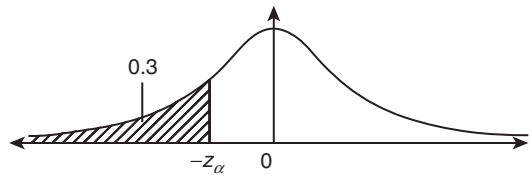
$$\therefore P(0 < z < z_\alpha) = 0.2$$

$$\therefore z_\alpha = 0.52$$

$$\text{(i.e.)} -0.52 = \frac{x - \mu}{\sigma}$$

$$-0.52 = \frac{40 - \mu}{\sigma}$$

$$\mu - 0.52\sigma = 40 \tag{5.17}$$



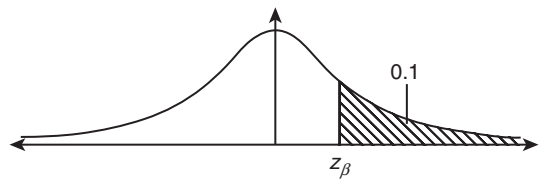
Given that student who passed with distinction are 10%, that is, who secured marks greater than or equal to 75.

$$P(x \geq 75) = 10\% = 0.1$$

$$P(z_\beta < z < \infty) = 0.1$$

$$\therefore P(0 < z < z_\beta) = 0.4$$

$$z_\beta = 1.28$$



Hence $z_\beta = \dfrac{x - \mu}{\sigma}$

$$1.28 = \frac{75 - \mu}{\sigma}$$

$$\mu + \sigma(1.28) = 75 \tag{5.18}$$

On solving equations (5.17) and (5.18), we get

$$\begin{array}{rrr} \cancel{\mu} + 1.28\sigma = 75 \\ \cancel{\mu} - 0.52\sigma = 40 \\ + \qquad - \\ \hline 1.8\sigma = 35 \\ \hline \sigma = 19.4 \end{array}$$

$$\sigma = 19.4$$

$$\mu = 75 - (1.28)\,(19.4)$$

$$= 50.1$$

Hence the average mark of the students is 50 marks.

EXAMPLE 5.10

Of a large group of men, 5% are under 60 inches in height and 40% are between 60 and 65 inches. Assuming normal distribution, find the mean and standard deviation.

Solution: Let x denote the height of men in inches.

Let μ and σ denote the mean height and standard deviation of the distribution.

Given that 5% are under 60 inches of height,

$$P(x < 60) = 5\% = 0.05$$

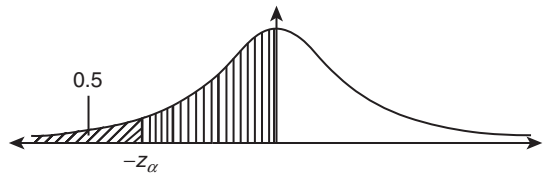
$$P(-\infty < z < -z_\alpha) = 0.05$$

$$P(z_\alpha < z < \infty) = 0.05$$

$$P(0 < z < \infty) - P(0 < z < z_\alpha) = 0.05$$

$$0.5 - P(0 < z < z_\alpha) = 0.05$$

$$\therefore P(0 < z < z_\alpha) = 0.45$$



From tables, $z_\alpha = 1.645$

$$\therefore -z_\alpha = \frac{x - \mu}{\sigma}$$

$$-1.645 = \frac{60 - \mu}{\sigma}$$

$$\mu - 1.645\sigma = 60 \tag{5.19}$$

Also given that 40% are between 60 and 65 inches.

From previous data

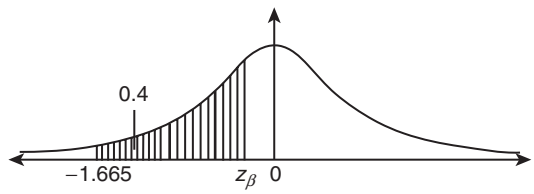
$$P(60 < x < 65) = 40\% = 0.4$$

$$P(-1.645 < z < z_\beta) = 0.4$$

$$P(z_\beta < z < 1.645) = 0.4$$

$$P(0 < z < z_\beta) = 0.05$$

$$z_\beta = 0.13$$



$$\therefore -z_\beta = \frac{x - \mu}{\sigma}$$

$$-0.13 = \frac{65 - \mu}{\sigma}$$

$$\mu - 0.13\sigma = 65 \tag{5.20}$$

On solving equations (5.19) and (5.20)

$$\mu - 1.645\sigma = 60$$

$$\mu - 0.13\sigma = 65$$

$$- \quad + \quad -$$

$$-1.515\sigma = -5$$

$$\sigma = 3.3$$

$$\mu = 60 + 1.645$$

$$= 65.45$$

Hence the average height of men is 65.45 inches with a standard deviation of 3.3 inches.

EXAMPLE 5.11

There are 600 business students in the post – graduate department of a university, and the probability for any student to need a copy of a text book from the university library on any day is 0.05. How many copies of the book should be kept in the university library, so that the probability may be 0.90 that none of the students needing a copy from the library has to come back disappointed.

Solution: Let n denote the number of students in the post graduate department of a university, p denote the probability that any student needs a copy of a book from the university library.

Given that $n = 600$, $p = 0.05$

Using binomial law with normal approximation we get

$\mu = nP$(mean of binomial distribution)

$= 600(0.05)$

$= 30$

$\sigma = \sqrt{npq}$ (Standard deviation of the binomial distribution)

$= \sqrt{600(0.05)(0.95)}$

$= 1 - 0.05$

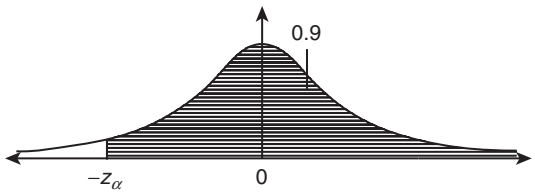
$= 0.95$

where $q = 1 - p = 5.34$

Let x denote the number of copies of text book required on any day.
Probability that more of the students come disappointed is greater than 0.9

$$p(x > x_\alpha) = 0.9$$

$$P(z > -z_\alpha) = 0.9$$



When

$$P(-\infty < z < z_\alpha) = 0.1$$

$$P(z_\alpha < z < \infty) = 0.1$$

$$P(0 < z < \infty) - P(0 < z < z_\alpha) = 0.1$$

$$P(0 < z < z_\alpha) = 0.4$$

$$z_\alpha = 1.28$$

$$-1.28 = \frac{x - \mu}{\sigma} = \frac{x - 30}{5.34}$$

$$\therefore x = 30 - (1.28)(5.34)$$

$$= 23.1 \cong 23$$

Hence the library should keep at least 23 copies of the book so that the probability is more than 90% that none of students has to come back disappointed from the library.

EXAMPLE 5.12

The probability that a patient recovers from a rare blood disease is 0.4. Of 100 people known to have contracted this disease, what is the probability that less than 30 survive.

Solution: Let x denote the number of patients that survive.

Using normal population to binomial variant,

$\mu = np$ (the mean of binomial variate)

Let n be the patients who were attacked by this disease, p be the probability that a patient recovers the disease.

Given $n = 100$, $p = 0.4$

$$\mu = 100(0.4) = 40$$

$$\sigma = \sqrt{npq} = \sqrt{100(0.4)(0.6)} = 4.899$$

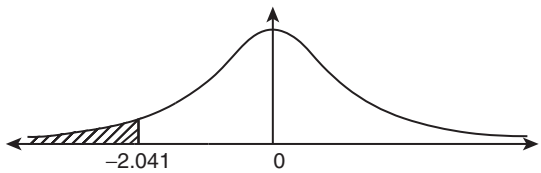
Probability that less than 30 survive from the disease

$$= P(x < 30)$$

Let $z = \dfrac{x - \mu}{\sigma}$

When $x = 30$, $z = \dfrac{30 - 40}{4.899}$,

$= -2.041$



$$P(x < 30) = P(z < -2.041)$$

$$= P(z > 2.041) = 0.5 - P(022 < 2.047)$$

$$= 0.5 - 0.4793$$

$$= 0.0207$$

$$\therefore P(z < -2.041) = 0.0207$$

EXAMPLE 5.13

Assume that 4% of the population over 65 years has Alzheimer's disease. Suppose a random sample of 9600 over 65 is taken. Find the probability p that tower than 400 of them have the disease.

Solution: This is a binomial experiment which has normal approximation as the sample size is very large.

Out of 9600, 4% has Alzheimer's disease.

Hence $p = 0.04$ $q = 0.96$ and $n = 9600$

Then the mean of the binomial experiment $\mu = np$

$$= 9600\ (0.04)$$

$$= 384$$

The standard deviation of binomial experiment $\sigma = \sqrt{npq}$

$$= \sqrt{(9600)(0.04)(0.96)} = 18.2$$

Let x be the number of people with Alzheimer's disease. Under the normal approximation, probabilities p that fever than 400 of them have the disease

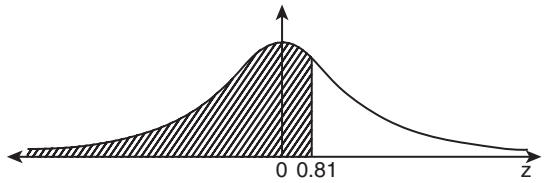
$= P(x < 400)$

$= P(z < 0.81)$

$= P(0 < z < 0) + P(0 < z < 0.81)$

$= 0.5 + 0.2892$

$= 0.7892$



$$z = \frac{x - \mu}{\sigma}$$

$z = \frac{400 - 384}{19.2}$, When $x = 400$

$= 0.81$

Work Book Exercises

1. Given a normal distribution with $\mu = 30$, $\sigma = 6$, find the normal curve area
   (i) To the right of $x = 17$,
   (ii) To the left of $x = 22$
   (iii) Between $x = 32$ and $x = 41$.
   (iv) The value of x that has 80% of the normal curve to the left
2. In a certain bakery the loaves have an average length of 30 cm and standard deviation of 2 cm. Assuming normal distribution for the lengths, What percentage of loaves are
   (i) Longer than 31.7 cm
   (ii) Between 29.3 and 33.5 cm in length
   (iii) Shorter than 25.5 cm?
3. A certain machine makes electrical resistors having a mean resistance of 40 ohms and standard deviation of 2 ohms. Assuming that resistance follows a normal distribution and can be measured to any degree of accuracy, what percentage of restores will have a resistance encoding 43 ohms?
4. Write the importance of normal distribution.
5. Prove that mean = mode = median for normal distribution.
6. If z is a normal variant, find
   (i) The area to the left of $z = -1.78$
   (ii) Area to the right of $z = 1.45$
   (iii) Area corresponding to $-0.80 \le z \le 1.53$
   (iv) Area to the left of $z = -2.52$ and to the right of $z = 1.83$
7. If x is normally distributed with mean 8 and S.D.H, find:
   (i) $P(5 \le x \le 10)$,
   (ii) $P(10 \le x \le 15)$,
   (iii) $P(x \le 15)$.
8. In a distribution exactly normal, 10.03% of the items are under 25 kilogram weight and 89.97 % of the items are under 70 kilogram weight. What are the mean and standard deviation of the distribution. A: $\mu = 47.5$ $\sigma = 17.578$ kg.

9. The loaves of rye bread distributed to local stores by a certain bakery have an average length of 30 cm and a standard deviation of 2 cm. Assuming that the lengths are normally distributed what percentages of the loaves are
   (i) Longer than 31.7 cm?
   (ii) between 29.3 and 33.5 cm in length,
   (iii) Shorter than 25.5 cm.
10. Assuming that height distribution of a group of men is normal, find the mean and standard deviation given that 84% of the men have height less than 65.2 inches and 68% have heights between 65.2 and 62.8 inches.

[**Ans.:** $\mu = 61.86$, $\sigma = 0.0397$]

11. If the mean and standard deviation of a normal distribution are 70 and 16 respectively, find $P(38 < x < 46)$. (JNTU, 2005S)

[**Ans.:** 0.044]

12. The average daily sale of 500 branch officials was ₹1,50,000 and the standard deviation was ₹15,000. Assuming the distribution to be normal, how many branches have sale between
   (i) ₹1,20,000 and 1,45,000
   (ii) ₹1,40,500 and ₹1,65,000
   (iii) more than ₹1,65000.

[**Ans.:** (i) 147 (ii) 295 (iii) 31 branches approximately]

13. If 20% of the memory chips made in a certain plant are defective, What are the probabilities that in a lot of 100 randomly chosen for inspection.
   (i) At most 15 will be defective
   (ii) Exactly 15 will be defectives

[**Ans.:** (i) 0.1292 (ii) 0.0454]

5.2 EXPONENTIAL DISTRIBUTION

The exponential distribution finds many applications in queuing theory and reliability problems. The time between arrivals at service facilities and time to failure of component parts and electrical systems are modeled by the exponential distribution.

Definition

A continuous random variable X assuming non negative values is said to have an exponential distribution, if its pdf is given by

$$f(x) = \begin{cases} \theta . e^{-\theta x}, & x \geq 0 \\ 0, & \text{otherwise} \end{cases} \tag{5.21}$$

Caution:

- The given function in (5.21) is a pdf since $\int_0^\infty \theta e^{-\theta x} dx = \theta \int_0^\infty e^{-\theta x} dx$

$$= \theta \frac{e^{-\theta x}}{-\theta} \Big|_0^\infty$$

$$= \theta \left[0 + \frac{1}{\theta} \right] = 1$$

$\therefore$ equation given in (5.21) is a pdf

- The cumulative distribution $F(x)$ is given by

$$F(x) = \int_0^x f(u)\,du = \theta \int_0^x e^{-\theta u}\,du$$

$$= \theta \frac{e^{-\theta u}}{-\theta}\Bigg|_0^x$$

$$F(x) = \begin{cases} 1 - e^{-\theta x}, & x \geq 0 \\ 0, & \text{otherwise} \end{cases}$$

- θ is called the parameter of exponential distribution
- X has an exponential distribution with parameter β has pdf also in the form

$$f(x, \beta) = \begin{cases} \frac{1}{\beta} e^{-x/\beta}, & x > 0 \\ 0, & \text{otherwise} \end{cases}$$

- The mean and variance for the above exponential distribution are $\mu = \beta$ and $\sigma^2 = \beta^2$.

Mean and Variance of Exponential Distribution

If X is exponential variate, then mean of $X = E(x) \int_0^\infty xf(x)dx$

$$E(x) = \int_0^\infty x\theta^{-\theta x}dx = \theta \int_0^\infty xe^{-\theta x}dx$$

$$= \theta \left[x \frac{e^{-\theta x}}{-\theta}\Bigg|_0^\infty + \frac{1}{\theta} \int_0^\infty e^{-\theta x}dx \right]$$

$$= \frac{\theta}{\theta} \frac{e^{-\theta x}}{-\theta}\Bigg|_0^\infty = \frac{1}{\theta}$$

$$\therefore \boxed{E(x) = \frac{1}{\theta}}$$

$$E(x^2) = \int_0^\infty x^2 f(x)\,dx = \theta \int_0^\infty x^2 e^{-\theta x}dx$$

$$= \theta x^2 \frac{e^{-\theta x}}{-0}\Bigg|_0^\infty + \frac{\theta}{\theta} \int_0^\infty e^{-\theta x}, 2x\,dx$$

$$= \frac{2}{\theta^2}$$

Variance of $x = E(x^2) - E(x^2)$

$$= \frac{2}{\theta} - \left(\frac{1}{\theta}\right)^2$$

$$= \frac{2}{\theta^2} = \left(\frac{1}{\theta}\right)^2 = \frac{1}{\theta^2}$$

MGF of Exponential Distribution

If X is an exponential variate, then

$$\text{MGF of } X = E(e^{tx})$$

$$\int_0^\infty e^{tx}(x)dx = \int_0^\infty e^{tx}\theta e^{-\theta x}dx$$

$$= \theta\int_0^\infty e^{-x(-t+8)}dx = \theta\frac{e^{-x(\theta-t)}}{\theta-t}\Bigg|_0^\infty$$

$$= \frac{\theta}{\theta-t}, \theta > t$$

$$= \theta(\theta-t)^{-1} = \frac{\theta}{\theta}\left(1-\frac{t}{\theta}\right)^{-1}$$

$$= 1+\frac{t}{\theta}+\frac{t^2}{\theta^2}+\cdots = \sum_{r=0}^{\infty}\left(\frac{t}{\theta}\right)^r$$

The moment about the origin can be obtained by the coefficient of the above expansion.

$$\mu_1' = E(x^r) = \text{coff of } \frac{t^r}{r!} \text{ in } M_x(t)$$

$$\mu_1' = \text{Mean} = \frac{1}{\theta}$$

$$\mu_2' = \frac{2}{\theta^2}$$

$$\therefore \text{ Variance } \mu_2' - (\mu_1')^2 = \frac{2}{\theta^2} - \frac{1}{\theta^2} = \frac{1}{\theta^2}$$

Memory Less Property of the Exponential Distribution

This is explained through an example. Suppose in case of an electronic component, where distribution of lifetime has an exponential distribution, the probability that the component lasts say t hours (ie) $P(x \geq t)$ is the same as the conditional probability $P(x \geq t_0 + \frac{t}{x} \geq t_0)$.

So if the component makes it to t_0 hours, the probability of lasting additional t hours is the same as the probability of lasting t hours.

Theorem: The exponential distribution lacks memory (ie) if X has an exponential distribution, then for every constant $a \geq 0$, one has $P\left(\frac{y \leq x}{x \geq a}\right) = P(x \leq x)$ for all x where $y = x - a$.

Proof: If X is an exponential variate, then the pdf of X is given by $f(x) = \theta e^{-\theta x}$; $\theta > 0$, $0 < x < \infty$.

The condition probability $P\left(\frac{y \leq x}{x \geq a}\right)$ is defined as

$$P\left(\frac{y \leq x}{x \geq a}\right) = \frac{P(y \leq x \cap x \geq a)}{P(x \geq a)}$$

Consider $P(y \leq x \cap x \geq a) = P(x - a \leq x \cap x \geq a)$

$$= P(x \leq a + x \cap x \geq a)$$

$$= P(a \le x \le a + x)$$

$$= \int_a^{a+x} f(x)dx = \int_a^{a+x} \theta e^{-\theta x} dx$$

$$= \theta \int_a^{a+x} e^{-\theta x} dx = \theta \frac{e^{-\theta x}}{-\theta}\Bigg|_a^{a+x} dx$$

$$\therefore p(y \le x \cap x \ge a) = e^{-a\theta}(1 - e^{-\theta x})$$

$$p(x \ge a) = \int_a^{\infty} f(x)dx = \theta \int_a^{\infty} e^{-\theta x} dx = \theta \frac{e^{-\theta x}}{-\theta}\Bigg|_a^{\infty}$$

$$= e^{-a\theta}$$

$$\therefore p\left(\frac{y \le x}{x \ge a}\right) = \frac{(1 - e^{-\theta x})e^{-a\theta}}{e^{-a\theta}}$$

$$= 1 - e^{-\theta x} \tag{5.22}$$

In addition, $P(x \le x) = \int_0^x \theta e^{-\theta x} dx = \theta \frac{e^{-\theta x}}{-\theta}\Big|_0^x$

$$= 1 - e^{-\theta x} \tag{5.23}$$

∴ From the above two equations, we observe that

$$P\left(\frac{y \le x}{x \ge a}\right) = P(x \le x)$$

Hence, exponential distribution lacks memory.

Worked Out Examples

EXAMPLE 5.14

If $x_1, x_2, \ldots x_n$ are independent random variables, where x_i follows exponential distribution with parameter θ_i; $i = 1, 2, \ldots n$; then show that $z = \min(x_1, x_2, \ldots x_n)$ has exponential distribution with parameter $\sum_{i=1}^{n} \theta_i$

Solution: Consider the function $G_{\bar{z}}(z) = P(Z < z)$

$$= 1 - P[\min(x_1, x_2, \cdots x_n) > z]$$

$$= 1 - P[x_i > z; i = 1, 2, \cdots n]$$

$$= 1 - \prod_{i=1}^{n} P(x_i > z) = 1 - \prod_{i=1}^{n} [1 - P(x_i \le z)]$$

$$= 1 - \prod_{i=1}^{n} 1 - F_{x_i}(z)$$

Where $F_{x_i}(z)$ is the distribution function of x_i

$$\therefore G_{z}(z) = 1 - \prod_{i=1}^{n} [1 - (1 - e^{-\theta iz})]$$

{$\because$the distribution function of exponential function is

$$F(x) = 1 - e^{-\theta x}\}$$

$$\therefore G_{z}(z) = 1 - \prod_{i=1}^{n} e^{-\theta iz}$$

$$= 1 - (e^{-\theta_1 z} e^{-\theta_2 z} \cdots e^{-\theta_n z})$$

$$= 1 - e^{(-\theta_1 + \theta_2 + \cdots + \theta_n) z}$$

$$= 1 - e^{-(-\theta_1, + \theta_2 + \cdots + \theta_n) z}$$

$$G_z(z) = \begin{cases} 1 - e - \left(\sum_{i=1}^{n} \theta_i \right) z, & z > 0 \\ 0, & \text{otherwise} \end{cases}$$

$\therefore$the pdf of z is given by,

$$g_z(z) = \begin{cases} \left(\sum_{i=1}^{n} \theta_i \right) e^{-\left(\sum_{i=1}^{n} \theta_i \right) z}, & z > 0 \\ 0, & \text{otherwise} \end{cases}$$

Hence $z = \min(x_1, x_2, \ldots x_n)$ is an exponential variant with parameter $\sum_{i=1}^{n} \theta_i$.

EXAMPLE 5.15

If X has an exponential distribution with mean 2, find $P\left[\left(\frac{x<1}{x<2}\right)\right]$

Solution: The pdf of exponential distribution with parameter θ, is given by $f(x) = \begin{cases} \theta e^{-\theta x}, & x \geq 0 \\ 0, & \text{Otherwise} \end{cases}$

$$P\left[\frac{(x<1)}{(x<2)}\right] = \frac{P[(x<1) \bigcap (x<2)]}{P(x<2)}$$

$$= \frac{P(x<1)}{P(x<2)}$$

Since the event common to both $(x < 1)$ and $(x < 2)$ is $(x < 1)$.

$$\therefore p\left[\left(\frac{x<1}{x<2}\right)\right] = \frac{\int_0^1 f(x)\,dx}{\int_0^2 f(x)\,dx}$$

$$= \frac{\int_0^1 \cancel{\theta} e^{-\theta x} dx}{\cancel{\theta} \int_0^2 e^{-\theta x} dx}$$

$$= \frac{\left. \frac{e^{-\theta x}}{-\theta} \right|_0^1}{\left. \frac{e^{-\theta x}}{-\theta} \right|_0^2}$$

$$= \frac{1 - e^{-\theta}}{1 - e^{-2\theta}}$$

EXAMPLE 5.16

If x and y are independent with common probability density function given by $f(x) = \begin{cases} e^{-x}, & x \geq 0 \\ 0, & x < 0 \end{cases}$
Find the pdf for $x - y$

Solution: Since x and y are independent and identically distributed random variables, their joint pdf is given by

$$f_{xy}(x, y) = \begin{cases} e^{-(x+y)}, & x > 0, y > 0 \\ 0, & \text{otherwise} \end{cases}$$

$$\text{Let } u = x - y \qquad \Rightarrow \qquad x = u + v$$
$$v = y \qquad\qquad\qquad\quad y = v$$

The Jacobian of the variables (x, y) w.r.t (u, v) is given by $J\left(\frac{x, y}{u, v}\right) = \begin{vmatrix} 1 & 1 \\ 1 & 1 \end{vmatrix} = 1$

Thus the joint pdf for the variables (u, v) is given by $g(u, v) = e^{-(u+v+v)}$, $-\infty < u < \infty$

$$x = u + v,\ u = x - v \Rightarrow v = x - u$$

Hence $v > -u$ if $-\infty < u < 0$ and $v > 0$ if $u > 0$
Thus for $-\infty < n < 0$; the marginal density function of u is given by

$$g(u) = \int_{-u}^{\infty} g(u, v) dv = \int_{-u}^{\infty} e^{-(u+2v)} dv$$

$$= e^{-u} \left. \frac{e^{-2v}}{-2} \right|_{-u}^{\infty} = \frac{1}{2} e^{u}$$

For $u > 0$ the marginal density function of u is given by

$$g(u) = \int_{-u}^{\infty} g(u, v) dv = \left. \frac{e^{-u} e^{-v}}{-2} \right|_{-u}^{\infty} = \frac{1}{2} e^{-u}$$

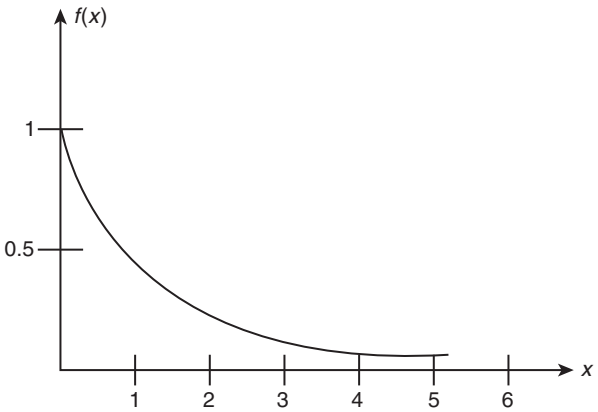
Hence the pdf of $u = x - y$ is given by

$$g(u) = \begin{cases} \frac{1}{2}e^{u}, & -\infty < u < 0 \\ \frac{1}{2}e^{-u}, & -\infty < u < 0 \end{cases}$$

Which means $g(u) = \frac{1}{2}e^{-|u|}, -\infty < u < \infty$

Caution: The above obtained density function is the pdf of standard Laplace distribution.

Graph of Exponential Distribution



Worked Out Examples

EXAMPLE 5.17

The length of time for one individual to be served at a cafeteria is a random variable having an exponential distribution with a mean of 4 minute. What is the probability that a person is served in less than 3 minutes on at least 4 of the next days?

Solution: The density function of exponential distribution is

$$f(x) = \theta e^{-\theta x}, x \geq 0$$

Where the mean of the exponential distribution $= \frac{1}{\theta}$

$$= \frac{1}{4}$$

Probability that a person is served in less than 3 minutes is

$$= P(x < 3) = \int_0^3 f(x)\,dx = \frac{1}{4}\int_0^3 e^{\frac{-x}{4}}\,dx$$

$$= \left.\frac{e^{\frac{-x}{4}}}{\frac{-1}{4}}\right|_0^3 = [1 - e^{\frac{-3}{4}}]$$

Let x represent number of days on which a person is served less than 3 minutes. $\therefore p = 1 - e^{\frac{-3}{4}}$, $q = 1 - p = e^{\frac{-3}{4}}$ Using binomial distribution,

Probability that it is served at least 4 days out of 6 days

$$= P(x = 4) + P(x = 5) + P(x = 6)$$
$$= 6C_4(1 - e^{\frac{-3}{4}})^4(e^{\frac{-3}{4}})^2 + 6C_5(1 - e^{\frac{-3}{4}})^5(e^{\frac{-3}{4}})^1 + 6C_6(1 - e^{\frac{-3}{4}})^6$$
$$= 0.3968$$

EXAMPLE 5.18

The life in year of a certain type of electrical switch has an exponential distribution with an average life of $\beta = 2$. If 100 of three switches are installed in different system, what is the probability that at most 30 fail during the first year?

Solution: Given that average life of electric switch (say X) is $\beta = 2$. Since it follows exponential distribution, Probability that electric switch fails during the first year is

$$P(x < 1) = \int_0^1 f(x)\,dx = \frac{1}{\beta}\int_0^1 e^{\frac{-x}{\beta}}\,dx = \frac{1}{2}\int_0^1 e^{\frac{-x}{2}}\,dx$$
$$= (1 - e^{\frac{-1}{2}})$$

Let X represent now the number of switches which fail during the first year. Using binomial distribution, $p = 1 - e^{\frac{-1}{2}}$, $q = e^{\frac{-1}{2}}$

Probability that atmost 30 switches fail during the first year out of 100 installed

$$= P(x \le 30) = \sum_{x=0}^{30} nC_x p^x q^{n-x}$$
$$= \sum_{x=0}^{30} 100C_x(1 - e^{\frac{-1}{2}})^x(e^{\frac{-1}{2}})^{100-x}$$

Work Book Exercise

14. (i) If X has exponential distribution with mean 2, find $P\left[\frac{(x < 1)}{(x < 2)}\right]$

    (ii) If $x \sim E(\lambda)$ with $P(x \le 1) = P(x > 1)$, find Var(x).

5.3 GAMMA DISTRIBUTION

The Gamma distribution derives its name from the well-known Gamma function which is studied in many area of mathematics.

The gamma function is defined by

$$\Gamma(n) = \int_0^\infty e^{-x}x^{n-1}dx \quad \text{for } n > 0$$

Some important properties of gamma function are as follows:

- $\Gamma(1) = 1$
- $\Gamma(n) = (n-1)!$
- $\Gamma(n+1) = n\Gamma(n)$
- $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$

Definition

A continuous random variable X which has following probability law

$$f(x) = \begin{cases} \dfrac{e^{-x}x^{\lambda-1}}{\Gamma(\lambda)}, & x > 0; \quad 0 < x < \infty \\ 0, & \text{otherwise} \end{cases} \tag{5.24}$$

is known as gamma variate with parameter λ, its distribution is called gamma distribution.

Caution:

- A variate follows Gamma distribution with parameter λ is written as $x \sim \Gamma(\lambda)$
- The function given in equation (5.24) is a probability density function since $\int_{-\infty}^{\infty} f(x)dx = \int_{0}^{\infty} \frac{e^{-x}x^{\lambda-1}}{\Gamma(\lambda)}dx = \frac{1}{\Gamma(\lambda)}\Gamma(\lambda) = 1.$
- A continuous random variable X having the following pdf is said to have Gamma distribution with two parameters a and λ:

$$f(x)\begin{cases} = \dfrac{a\lambda}{\Gamma(\lambda)}e^{-ax}x^{\lambda-1}, & a > 0, \lambda > 0, 0 < x < \infty \\ = 0, & \text{otherwise} \end{cases}$$

- The cumulative distribution function for Gamma variate is

$$F(x) = \int_0^x f(u)du \begin{cases} = \dfrac{1}{\Gamma(x)}\displaystyle\int_0^x e^{-u}u^{\lambda-1}du \; x < 0 \\ = 0 \quad \text{otherwise} \end{cases}$$

- This is called Incomplete Gamma function.

Mean and Variance of Gamma Distribution

If X is a Gamma variate, then

$$\text{Mean of } X = \mathrm{E}(X) = \int_0^{\infty} xf(x)dx$$

$$= \int_0^{\infty} x\,\frac{e^{-x}x^{\lambda-1}}{\Gamma(\lambda)}dx$$

$$= \frac{1}{\Gamma(\lambda)}\int_0^{\infty} x^{\lambda}e^{-x}dx$$

$$= \frac{1}{\Gamma(\lambda)}\Gamma(\lambda+1)$$

$$= \frac{\lambda\Gamma(\lambda)}{\Gamma(\lambda)} = \lambda$$

$$\boxed{E(X) = \lambda}$$

$$E(X^2) = \int_0^\infty x^2 f(x)\,dx$$

$$= \int_0^\infty x^2 \frac{e^{-x}x^{\lambda-1}}{\Gamma(\lambda)}dx$$

$$= \frac{1}{\Gamma(\lambda)}\int_0^\infty e^{-x}x^{\lambda+1} \quad dx$$

$$= \frac{\Gamma(\lambda+2)}{\Gamma(\lambda)} = \frac{\lambda(\lambda+1)\Gamma(\lambda)}{\Gamma(\lambda)}$$

$$= \lambda(\lambda+1)$$

Variance of $X = V(X) = \mathrm{E}(X^2) - [\mathrm{E}(X)]^2$

$$= \lambda(\lambda+1) - \lambda^2$$

$$V(X) = \lambda$$

Caution:

- Mean and variance of Gamma distribution coincide = λ, the parameter of distribution.
- Mean variance of Gamma distribution with two parameters is, mean = $\frac{\lambda}{a}$ and variance = $\frac{\lambda}{a^2}$

MGF of Gamma Distribution

If X is a Gamma variant then,

$$\text{MGF of } X = E[e^{tx}] = \int_0^\infty e^{tx}f(x)dx = \int_0^\infty e^{tx}\frac{e^{-x}x^{\lambda-1}}{\Gamma(x)}dx$$

$$= \frac{1}{\Gamma(\lambda)}\int_0^\infty e^{-(1-t)x}x^{\lambda-1}dx$$

We have properties of Gamma functions as follows:

$$\int_0^\infty x^m e^{-ax^n}dx = \frac{1}{\frac{m+1}{n}}\Gamma\left(\frac{m+1}{n}\right)$$

where m and n are positive constants. Here $m = \lambda - 1$, $n = 1$, $a = (1 - t)$

$$\therefore \int_0^\infty e^{-(1-t)x}x^{\lambda-1}dx = \frac{1}{\frac{(\lambda-1+1)}{(1-t)}}\Gamma\left(\frac{\lambda-1+1}{1}\right)$$

$$= \frac{1}{(1-t)^\lambda}\Gamma(\lambda)$$

Substituting this in the above MGF, we get

$$m_x(t) = \frac{1}{\Gamma(\lambda)} \frac{\Gamma(\lambda)}{(1-t)\lambda}, |t| < 1$$
$$= (1-t)^{-\lambda}, |t| < 1$$
$$= \lambda C_0 + \lambda C_1 t + \lambda C_2 t^2 + \cdots + \lambda C_r t^r + \cdots$$

The Coefficient of $\frac{t^r}{r!}$ will give different moments about the origin.

$$\mu_r' = \lambda C_r \, r!$$
$$\mu_1' = \lambda C_1 = \lambda$$
$$\mu_2' = \lambda C_2 \, 2! = \frac{\lambda(\lambda-1)}{2!} 2! = \lambda(\lambda-1)$$
$$\text{var}(x) = \mu_2' - (\mu_1')^2 = \lambda(\lambda-1) - \lambda^2$$
$$= \lambda^2 - \lambda - \lambda^2 = -\lambda$$

Worked Out Examples

EXAMPLE 5.19

In a certain city, the daily consumption of water (in millions of litres) follows approximately a Gamma distribution with $\lambda = 2$ and $a = \frac{1}{3}$ of the daily capacity of that city is 9 million liters of water, what is the probability that on any given day the water supply is inadequate?

Solution: The density function of Gamma distribution with two parameters

$$f(x) = \frac{a\lambda}{\Gamma(\lambda)} e^{-ax} x^{\lambda-1}, \; a > 0, \lambda > 0, 0 < x < \infty$$
$$= \frac{1}{3^2 \Gamma(2)} e^{\frac{-x}{3}} x^{2-1}$$
$$= \frac{1}{9} e^{\frac{-x}{3}} x$$

Let X denote the supply of water in million liters. Probability that water supply is inadequate on any day (i.e) water supply is less than the capacity = $P(X < 9)$

$$= \int_0^9 f(x)dx = \frac{1}{9} \int_0^9 x e^{\frac{-x}{3}} dx$$
$$= \frac{1}{9} \left[\frac{x e^{\frac{-x}{3}}}{\frac{-1}{3}} \Bigg|_0^9 - \int_0^9 \frac{e^{\frac{-x}{3}}}{\frac{-1}{3}} dx \right] = \frac{1}{3}(-9e^{-3}) + \frac{1}{3} \frac{e^{\frac{-x}{3}}}{\frac{-1}{3}} \Bigg|_0^9$$
$$= \frac{1}{3}[-9e^{-3} + 1 - e^{-3}] = \frac{1}{3}(1 - 10e^{-3})$$

EXAMPLE 5.20

In a certain city the daily consumption of electric power in millions of kilowatt hours is a random variable X having a gamma distribution with mean $\mu = 6$ and variance $\sigma^2 = 12$.

(i) Find the values of a and λ.

(ii) Find the probability that on any given day, the daily power consumption will exceed 12 million kilowatt hours.

Solution:

(i) The density function of Gamma distribution is

$$f(x) = \frac{a^{\lambda}}{\sigma(\lambda)} e^{-ax} x^{\lambda-1}$$

The mean and variance of Gamma distribution are

$$\mu = \frac{\lambda}{a}, \text{ variance } \sigma^2 = \frac{\lambda}{a^2}$$

Given $\mu = 6$ and $\sigma^2 = 12$

$$\frac{\lambda}{a} = 6, \frac{\lambda}{a^2} = 12 \Rightarrow \frac{\frac{\lambda}{a}}{\frac{\lambda}{a^2}} = \frac{6}{12} \Rightarrow a = \frac{1}{2}, \lambda = 3$$

(ii) Let X be daily power consumption in the city probability that daily power consumption will exceed 12 million kilowatt hours is

$$P(x > 12) = 1 - P(x < 12) = 1 - \int_0^{12} f(x) fx$$

$$= 1 - \int_0^{12} \frac{1}{2^3 \Gamma(3)} e^{\frac{-x}{2}} x^{3-1} dx$$

$$= 1 - \frac{1}{16} \int_0^{12} x^2 e^{\frac{-x}{2}} dx = 1 - \frac{1}{16} \left[x^2 \frac{e^{\frac{-x}{2}}}{\frac{-1}{2}} \Big|_0^{12} + \int_0^{12} e^{\frac{-x}{2}} \cdot 2x dx \right]$$

$$= 1 - \frac{1}{16} \left[-2(144)e^{-6} + 4\, x \frac{e^{\frac{-x}{2}}}{\frac{-1}{2}} \Big|_0^{12} + 8 \int_0^{12} e^{\frac{-x}{2}} dx \right]$$

$$= 1 - \frac{1}{16} \left[288\, e^{-6} - 96e^{-6} + 8 \frac{e^{\frac{-x}{2}}}{\frac{-1}{2}} \right]^0 = 1 - \frac{101}{8} e^{-6}$$

5.4 WEIBULL DISTRIBUTION

This distribution deals with problems, where identical components are subjected to identical environment conditions and fail at different times that are unpredictable.

Examples where this distribution can be applied are when a fuse burns out, when a steel column buckles etc.

Definition

A continuous random variable X has Weibull distribution with parameters α and β if its density function is given by

$$f(x) = \begin{cases} \alpha\beta x^{\beta-1} e^{-\alpha x^{\beta}}, & x > 0 \\ 0, & \text{otherwise} \end{cases} \tag{5.25}$$

When $\alpha > 0$ and $\beta > 0$

Caution:

The probability function given in (5.25) is a pdf since $\int_0^{\infty} \alpha\beta\, x^{\beta-1} e^{-\alpha x^{\beta}} dx$

$$x^{\beta} = y \qquad \therefore = \alpha \int_0^{\infty} e^{-\alpha y} dy$$

$$\beta x^{\beta-1} dx = dy \qquad = \alpha \frac{e^{-\alpha y}}{-2}\Bigg|_0^{\infty} = 1$$

$\therefore$ The probability function given is a pdf.

Mean and Variance of Weibull Distribution

If X is a Weibull variate, then

$$\text{Mean of } X = E(x) = \int_0^{\infty} x f(x) dx$$

$$= \beta \int_0^{\infty} x . x^{\beta-1} e^{-2x^{\beta}} dy$$

$$x\beta = y \Rightarrow x = y^{\frac{1}{\beta}}$$

$$\beta x^{\beta-1} dx = dy$$

$$\therefore E(x) = \alpha\beta \int_0^{\infty} y^{\frac{1}{\beta}} e^{-\alpha y} dy$$

$$\therefore E(x) = \alpha\beta \int_0^{\infty} y^{\frac{1}{\beta}} e^{-\alpha y} dy$$

We have from the property of Gamma function,

$$\int_0^{\infty} x^m e^{-\alpha x^n} dx = \frac{\Gamma\left(\frac{m+1}{n}\right)}{n a^{\frac{(m+1)}{n}}} \tag{5.26}$$

$$m = \frac{1}{\beta}, \quad a = \alpha, \quad n = 1$$

$$\int_{\alpha}^{\infty} y^{\frac{1}{\beta}} e^{-\alpha y} dy = \frac{\Gamma\left(\frac{1}{\beta}+1\right)}{1 \cdot \alpha^{\left(\frac{1}{\beta}+1\right)}}$$

$$\therefore E(x) = \frac{\alpha\beta' \cdot \Gamma\left(\frac{1}{\beta}+1\right)}{\beta\alpha^{\frac{1}{\beta}+1}} = \alpha^{\frac{-1}{\beta}}\Gamma\left(\frac{1}{\beta}+1\right)$$

$$E(x^2) = \int_0^\infty x^2 f(x)dx = \int_0^\infty x^2 \alpha\beta x^{\beta-1} e^{-\alpha x^\beta} dx$$

$$= \alpha\beta \int_0^\infty x^2 e^{-\alpha x^\beta} (x^{\beta-1}dx)$$

$$= \alpha\beta \int_0^\infty y^{\frac{2}{\beta}} e^{-\alpha y} \frac{dy}{\beta}$$

Let $x = y$

$$\beta x \beta^{-1} dx = dy$$

$$= \alpha\beta \int_0^\infty y^{\frac{2}{\beta}} e^{-\alpha y} \frac{dy}{\beta}$$

From equation (5.26),

$$\int_0^\infty y^{\frac{2}{\beta}} e^{-\alpha y} dy = \frac{\Gamma\left(\frac{2}{\beta}+1\right)}{1 \cdot \alpha^{\frac{2}{\beta}+1}}$$

$$\therefore E(x^2) = \alpha\beta \frac{\Gamma\left(\frac{2}{\beta}+1\right)}{\alpha^{\frac{2}{\beta}} \cdot \alpha} = \frac{\Gamma\left(\frac{2}{\beta}+1\right)}{\alpha^{\frac{2}{\beta}}}$$

$$\therefore \text{Variance of } X = E(X^2) - [\mathrm{E}(X)]^2 = \frac{\Gamma\left(\frac{2}{\beta}+1\right)}{\alpha^{\frac{2}{\beta}}} - \frac{\Gamma\left(\frac{1}{\beta}+1\right)^2}{\alpha^{\frac{2}{\beta}}}$$

$$= \frac{1}{\alpha^{\frac{2}{\beta}}}\left[\Gamma\left(1+\frac{2}{\beta}\right) - \Gamma\left(1+\frac{1}{\beta}\right)^2\right]$$

Caution:

- α and β are the parameters of Weibull distribution.
- A variable X follows Weibull distribution is written as $X \sim W(x, \alpha, \beta)$.

Worked Out Problem

EXAMPLE 5.21

Suppose the lifetime of a certain kind of an emergency backup battery (in hours) is a random variable X having the Weibull distribution with $\alpha = 0.1$ and $\beta = 0.5$. Find:

(i) The mean lifetime of these batteries

(ii) The probability that such a battery will last more than 300 hours?

Solution:

(i) The mean of Weibull distribution is given by $E(x) = \alpha^{\frac{1}{\beta}}\Gamma\left(1+\frac{1}{\beta}\right)$

$$= (0.1)^{-2}\Gamma(3) = 100(2!)$$

The mean lifetime of these batteries = 200 hours.

(ii) Probability that battery will last more than 300 hours = $P(X > 300)$

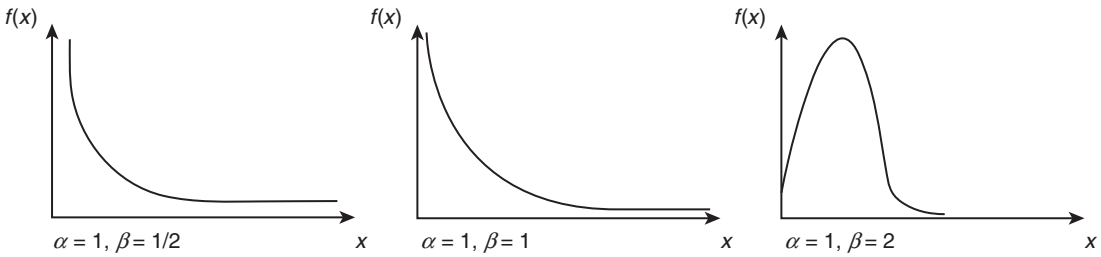
$$= \int_{300}^{\infty} \alpha\beta e^{-\alpha x^{\beta}} x^{\beta-1} dx$$

$$= \int_{300}^{\infty} (0.05)\, x^{-0.5} e^{-0.1(0.5)} dx$$

$$= e^{-0.1(300)^{0.5}}$$

$$= 0.177$$

Graph of Weibull Distribution



Cumulative Distribution of Weibull Distribution

If X is Weibull variate, then the cumulative distribution of x is given by $F(x) = 1 - e^{-\alpha x^{\beta}}$ for $x \geq 0$.

Since the cumulative distribution of x is defined as

$$F(x) = \int_0^x f(x)dx$$

$$= \int_0^x \alpha\beta x^{\beta-1} e^{-\alpha x^{\beta}} dx \quad \text{Let } x^{\beta} = y$$

$$\beta x^{\beta-1} dx = dy$$

$$= \int \alpha\beta \frac{1}{\beta} e^{-\alpha\beta} dy$$

$$= \alpha \int e^{-2y} dy = \alpha\left[\frac{e^{-dy}}{-\alpha}\right]_0^x = \frac{\alpha}{\alpha}\left[1 - e^{-2x\beta}\right]$$

$$= 1 - e^{-2x\boldsymbol{\beta}}$$

Failure Rate for Weibull Distribution

The Weibull distribution is also applied to reliability and life-testing problems like Gamma and Exponential distributions, such as "the time to failure" or "life length" of a component, which is measured from some specified time until it fails. Let us represent this time to failure by the continuous random variable T with pdf, $f(t)$ where $f(t)$ is the Weibull distribution.

The failure rate for the Weibull distribution at time 't' is given by

$$z(t) = \frac{f(t)}{e^{-\alpha t^{\beta}}}, t > 0$$

$$z(t) = \alpha\beta t^{\beta-1}, t > 0$$

Some points which can be noted here are as follows:

(i) If $\beta = 1$, the failure rate = α, a constant. Here in this case the lack of memory prevails.
(ii) If $\beta > 1$, $z(t)$ is an increasing function of t which indicates the components wears over time.
(iii) If $\beta < 1$, $z(t)$ is a decreasing function of time and hence the components strengthens over time.

Work Book Exercises

15. Find the mean and variance of the daily water consumption if in a certain city, the daily consumption of water (in millions of water) follows a Gamma distribution with 2 = 2 and $\beta = 3$, If the daily capacity of that city is 9 millions of litres of water.
16. In a biomedical research activity it was determined that the survival time in weeks of an animal when subjected to a certain type exposure of gamma radiation has a gamma distribution with $\alpha =$ 5 and $\beta = 10$.
    (i) What is the mean survival time of a randomly selected animal of the type used in the experiment
    (ii) What is the variance of survival time
    (iii) What is the probability that an animal survives more than 30 weeks.
17. Derive the mean and variance of the Weibul distribution.
18. Suppose that the life time of a certain kind of an emergency backup battery (in hours) is a random variable x having Weibull distribution with $\alpha = 0.1$ and $\beta = 0.5$,
    (i) Find mean lifetime of these batteries
    (ii) Find the probability that such a battery will last more than 300 hours.
19. The amount of time that a surveillance camera will run without having to be reset is a random variable passing the experimental distribution with $\beta = 60$ days. Find the probability that such a camera:
    (i) Have to be reset in less than 30 days
    (ii) Not have to be reset in atleast 50 days.

5.5 CENTRAL LIMIT THEOREM

Generally it is not possible to determine any distribution exactly without actual knowledge of the population. However, it is possible to the limiting distribution as $n \to \infty$ of a random variable whose values are

closely related to $\bar{x}$. Here we assume that the population has a finite variance σ^2. This random variable that is referred to is called standardized sample mean.

Theorem: If $\bar{x}$ is the mean of a sample of size n taken from a population having the mean μ and the finite variance σ^2, then

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

is a random variable whose distribution function approaches normal distribution of $n \to \infty$.

Caution:

- The distribution of $\bar{x}$ is approximately normal with mean μ and variance $\frac{\sigma^2}{n}$ whenever n is large.
- The normal distribution provides an approximation to the sampling distribution of the mean $\bar{x}$ for n as small as 25 or 30.
- If the random samples come from a normal population, the sampling distribution of the mean is normal, regardless of the size of the sample.

Worked Out Examples

EXAMPLE 5.22

A random sample of size 100 is taken from an infinite population having the mean $\mu = 76$ and variance $\sigma^2 = 256$. What is the probability that $\bar{x}$ will be between 75 and 78?

Solution:

Given that sample size $n = 100$

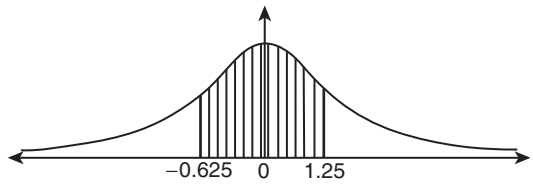
Mean $\mu = 76$, variance $\sigma^2 = 256$

Probability that $\bar{x}$ will be between 75 and 78 is $P(75 < \bar{x} < 78)$

Using the central limit theorem, $z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$

When $\bar{x} = 75$, $z_1 = \frac{75 - 76}{\sqrt{\frac{256}{100}}} = -0.625$

When $\bar{x} = 78$, $z_2 = \frac{78 - 76}{\sqrt{\frac{256}{100}}} = 1.25$



$$\therefore P(75 < \bar{x} < 78) = P(-0.625 < z < 1.25)$$
$$= P(-0.625 < z < 0) + P(0 < z < 1.25)$$
$$= P(0 < z < 0.625) + P(0 < z < 1.25)$$
$$= 0.234 + 0.3944 = 0.628$$

EXAMPLE 5.23

A random sample of size 100 is taken from an infinite population having the mean 30 and standard deviation 20. What is the

(i) Probability that $\bar{x}$ will be greater than 85?

(ii) Probability that $\bar{x}$ will lie between 75 and 85?

Solution: Given that sample size = 100

Mean of the sample, $\bar{x} = 80$

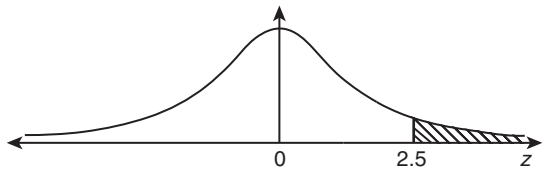
Standard deviation, $\sigma = 20$

The standard normal variant, $z = \dfrac{\bar{x} - \mu}{\dfrac{\sigma}{\sqrt{n}}}$

(i) Probability that $\bar{x}$ will be greater than 85

$= P(\bar{x} > 85)$,

When $\bar{x} = 85$, $z = \dfrac{85 - 80}{\dfrac{20}{\sqrt{100}}} = \dfrac{5}{2} = 2.5$



$P(\bar{x} > 85) = P(z > 2.5)$

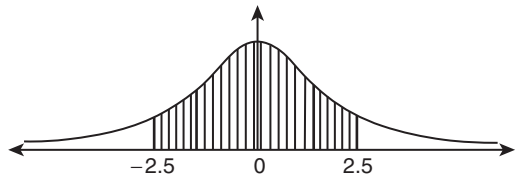
$= P(0 < z < \infty) - P(0 < z < 25)$

$= 0.5 - 0.498$

$= 0.0065$

(ii) Probability that $\bar{x}$ will be between 75 and 85 $= P(75 < \bar{x} < 85)$

When $\bar{x} = 75$, $z = \dfrac{75 - 80}{\dfrac{20}{\sqrt{100}}} = \dfrac{-5}{2} = -2.5$



$P(75 < \bar{x} < 85) = P(-2.5 < \bar{x} < 2.5)$

$= 2 \times P(0 < \bar{x} < 2.5)$

$= 2 \times 0.4938 = 0.\ 9876.$

EXAMPLE 5.24

The mean height of students of a class is 155 cms and standard deviation is 15. What is the probability that the mean height of 36 students is less than 157 cm?

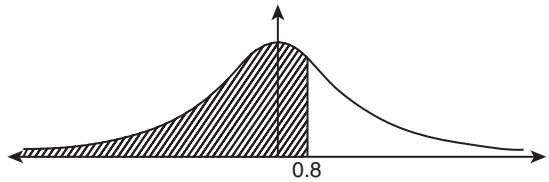
Solution: Mean height of students $\mu = 155$ cm

Standard deviations, $\sigma = 15$

Probability that mean height of 36 students is less than 157 cm $= P(\bar{x} < 157)$

When $\bar{x} = 157$, $z = \dfrac{\bar{x} - \mu}{\dfrac{\sigma}{\sqrt{n}}} = \dfrac{157 - 155}{\dfrac{15}{\sqrt{36}}} = \dfrac{+2}{2.5} = 0.8$

$$P(\bar{x} < 157) = p\,(z < 0.8)$$
$$= P(-\infty < z < 0) + P(0 < z < 0.8)$$
$$= 0.5 + 0.2881$$
$$= 0.7881$$



EXAMPLE 5.25

An electrical firm manufactures light bulbs that have a length of life that is approximately normally distributed with mean equal to 800 hours and a standard deviation of 40 hours. Find the probability that a random sample of 16 bulbs will have an average life of

(i) less than 775 hours?

(ii) Will be between 775 and 825 hours.

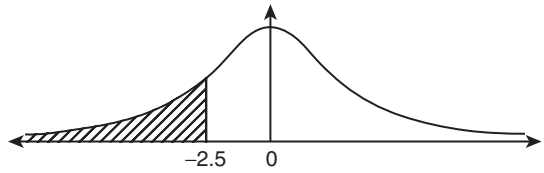
Solution: Mean life of light bulbs = 800 hours

Standard deviation of light bulbs $\sigma = 40$ hours

Probability that bulbs have an average life of less than 775 hours = $P(\bar{x} < 775)$

(i) When $\bar{x} = 775$, $z = \dfrac{\bar{x} - \mu}{\dfrac{\sigma}{\sqrt{n}}} = \dfrac{775 - 800}{\dfrac{40}{\sqrt{16}}} = -2.5$

$$P(\bar{x} < 775) = P(z < -2.5)$$
$$= P(Z > 2.5)$$
$$= -P(0 < z < \infty)$$
$$-P(0 < z < 2.5)$$
$$= 0.5 - 0.4938 = 0.0062$$

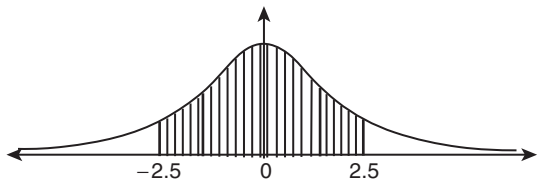


(ii) Probability that average life will be between 775 and 825 hours = $P(775 < \bar{x} < 825)$

When $\bar{x} = 825$,

$$z = \frac{\bar{x} - \mu}{\dfrac{\sigma}{\sqrt{n}}} = \frac{825 - 800}{\dfrac{40}{\sqrt{16}}} = 2.5$$

$$P(775 < \bar{x} < 825) = P(-2.5 < \bar{x} < 2.5)$$
$$= 2P(0 < z < 2.5)$$
$$= 2(0.4932) = 0.9876$$



EXAMPLE 5.26

The average life of a bread-making machine is 7 years with a standard deviation of 1 year. Assuming that the lives of these machines follow approximately normal distribution, find:

(i) The probability that mean life of a random sample of such machines falls between 6.4 and 7.2 years.

(ii) The value of $\bar{x}$ to the right of which 15% of the means computed from random samples of size 9 would fall.

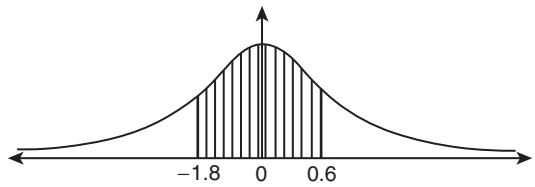
Solution: Given average life of machine = 7 years

Standard deviation of machine = 1 year

(i) Probability that mean life of random sample falls between 6.4 and 7.2 years
$= P(6.4 < \bar{x} < 7.2)$

When $\bar{x} = 6.4,\ t = \dfrac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} = \dfrac{6.4 - 7}{\frac{1}{\sqrt{9}}} = -1.8$

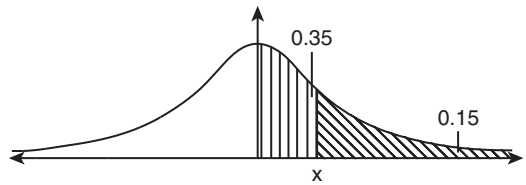
When $\bar{x} = 7.2,\ t = \dfrac{7.2 - 7}{\frac{1}{\sqrt{9}}} = 0.6$



$$P(6.4 < x < 7.2) = P(-1.8 < t < 0.6)$$
$$= P(0 < t < 1.8) + P(0 < t < 0.6)$$
$$= 0.45 + 0.225$$
$$= 0.675$$

(ii) Value of t to the height of which 15% means computed

$$0.15 = P(\bar{x} > t_1)$$



$$P(0 < t < t_1) = 0.35 \Rightarrow$$
$$t_1 = \frac{0.889 + 1.4}{2} = 1.1445$$
$$t_1 = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} \Rightarrow 1.445 = \frac{\bar{x} - 7}{\frac{1}{\sqrt{9}}} \Rightarrow \bar{x} = \frac{1.445}{3} + 7$$
$$\Rightarrow \bar{x} = 7.3815$$

20. Determine the probability that the sample mean are covered by a sample of 40 of 1 litre paint boxes will be between 510 to 520 square feet given that 1 litre of such paint box covers on the average 513.3 square feet with a standard deviation of 31.5 square feet.

[**Ans.:** 0.6552]

21. The amount of time that a bank clerk spends with a customer is a random variable with mean $\mu =$ 3.2 minutes and a standard deviation σ = 1.6 minutes. If a random sample of 64 customers is observed, find the probability that their mean time at the teller's counter is
    (i) At most 2.7 minutes
    (ii) More than 3.5 minutes
    (iii) At least 3.2 minutes and less than 3.4 minutes

DEFINITIONS AT A GLANCE

Normal Distribution: A symmetrical distribution which is symmetrical about the mean, bell shaped becomes sparse at the extremes and x-axis is a tangent at infinity.

Standard Normal Distribution: A normal probability distribution, has its mean zero and the standard deviation as 1.

Asymptote: It is a tangent to the curve at infinity. A line which appears to touch the curve but never touches it.

Exponential Distribution: Special Gamma distribution for $\alpha = 1$ and a distribution whose mean and standard deviation are the same.

Gamma Distribution: The distribution which uses Gamma function in its density function.

Weibull Distribution: A distribution which becomes bell-shaped for $\beta > 1$ and resemble normal curve and for $\beta = 1$ reduces to exponential distribution.

Points of Inflection: The points on the curve where the curve crosses the tangent.

FORMULAE AT A GLANCE

- The probability density function of Normal distribution is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{X-\mu}{\sigma}\right)^2}, \quad -\infty < x < \infty, -\infty < \mu < \infty, \sigma > 0$$

- The standard normal variate is defined as $z = \frac{x-\mu}{\sigma}$ whose mean is 0 and variance is1.
- The pdf of standard normal distribution is

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}}$$

- The mean, mode and median of a normal distribution coincide.
- Points of inflection of normal distribution are $x = \mu \pm \sigma$.
- Mean deviation about the mean of normal distribution is approximately $\frac{4}{5}\sigma$.
- Area property of Normal distribution:

$$P(\mu - \sigma < x < \mu + \sigma) = P(-1 < z < 1) = 0.6826$$
$$P(\mu - 2\sigma < x < \mu + 2\sigma) = P(-2 < z < 2) = 0.9544$$
$$P(\mu - 3\sigma < x < \mu + 3\sigma) = P(-3 < z < 3) = 0.9973$$

- MGF of normal distribution is $e^{\mu t + \frac{\sigma^2 t^2}{2}}$.
- MGF of standard normal variate is $e^{\frac{t^2}{2}}$.
- The ordered moments of normal variate vanish whereas the even ordered moments are given by,

$$\mu_{2r} = \sigma^{2r}[1 \cdot 3 \cdot 5 \cdots (2r-5)(2r-3)(2r-1)]$$

- Stirling's approximation to $n!$ for large n is

$$\underset{n\to\infty}{Lt}\, n! \sim \sqrt{2\pi} e^{-n} n^{n+\left(\frac{1}{2}\right)}$$

- In fitting normal distribution we can find

$$\mu = \frac{\Sigma f_i x_i}{\Sigma f_i} \qquad \text{and} \qquad \sigma = \sqrt{\frac{\Sigma f_i (x_i - \mu)^2}{\Sigma f_i}}$$

- The pdf of an exponential distribution is

$$f(x) = \theta e^{-\theta x}, \qquad x \geq 0$$

- Mean and variance of exponential distribution are $E(X) = \frac{1}{\theta}$ and $V(X) = \frac{1}{\theta^2}$ where θ is parameter.
- MGF of exponential distribution is $\sum_{r=0}^{\infty} \left(\frac{t}{\theta}\right)^r$.
- The exponential distribution lacks memory when $P\left(\frac{y \leq x}{x \geq a}\right) = P(x \leq x)$
- The pdf of Laplace distribution is

$$f(x) = \frac{1}{2} e^{-|x|}, \qquad -\infty < x < \infty$$

- The pdf of Gamma distribution is

$$f(x) = \frac{e^{-x} x^{\lambda - 1}}{\Gamma(\lambda)}, \quad \lambda > 0, 0 < x < \infty$$

- The pdf of Gamma variate with two parameters a and λ is,

$$f(x) = \frac{a^{\lambda}}{\Gamma(\lambda)} e^{-\alpha x} x^{\lambda - 1}, \qquad a > 0, \lambda > 0, 0 < x < \infty$$

- Mean of Gamma variate is λ and variance of Gamma variate is λ.
- MGF of Gamma distribution is $(1 - t)^{-\lambda}$, $|t| < 1$.
- The pdf of Weibull distribution is

$$f(x) = \alpha \beta x^{\beta - 1} e^{-\alpha x^{\beta}}, \qquad \alpha > 0, \ \beta > 0, x > 0$$

- The mean of Weibull distribution is

$$E(X) = \sigma^{-\frac{1}{\beta}} \Gamma\left(\frac{1}{\beta} + 1\right).$$

- Variance of Weibull distribution is

$$V(X) = \frac{1}{\alpha^{\frac{2}{\beta}}} \left[\Gamma\left(1 + \frac{2}{\beta}\right) - \Gamma\left(1 + \frac{1}{\beta}\right)^2 \right].$$

- The cumulative distribution of Weibull distribution is

$$F(x) = 1 - e^{-\alpha x^{\beta}}, \quad x \geq 0.$$

- The failure rate of Weibull distribution is given by

$$z(t) = \alpha\beta t^{\beta-1},\ t > 0$$

- Central limit theorem: If $\bar{x}$ is mean of a sample of size n, then $z = \dfrac{\bar{x} - \mu}{\dfrac{\sigma}{\sqrt{n}}}$ is a standard normal variate.

OBJECTIVE TYPE QUESTIONS

1. The distribution in which mean, median and mode coincide is _____.
   (a) Normal (b) Poisson
   (c) Exponential (d) Binomial

2. Suppose X is a normal variate with mean 10 and variance 4, then $P(X < 11) =$ _____.
   (a) 0.9915 (b) 0.8915
   (c) 0.6915 (d) 0.5

3. The probability density function for a standard normal variate is _____.
   (a) $f(z) = e^{\frac{-z^2}{2}}$ (b) $f(z) = \dfrac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}}$
   (c) $f(z) = \dfrac{1}{\sqrt{2\pi}} e^{-z^2}$ (d) $f(z) = \dfrac{1}{\sqrt{2\pi}} e^{z^2}$

4. The area under the normal curve between $\mu - 3\sigma$ and $\mu + 3 - \sigma$ is _____.
   (a) 0.9973 (b) 0.9544
   (c) 0.6826 (d) 0.5

5. The mean deviation from the mean for Normal distribution is _____ times the standard deviation.
   (a) $\dfrac{2}{5}$ (b) $\dfrac{1}{5}$
   (c) $\dfrac{3}{5}$ (d) $\dfrac{4}{5}$

6. If X is a normal variate with mean 30 and standard deviation 5, Then $P(X \geq 45) =$ _____.
   (a) 0.135 (b) 0.00135
   (c) 0.000135 (d) none

7. The normal curve is symmetric about _____.
   (a) x-axis (b) y-axis
   (c) z-axis (d) No symmetry

ANSWERS

1. (a) 2. (c) 3. (b) 4. (a) 5. (d) 6. (b) 7. (b)

6 Sampling Theory and Distribution

Prerequisites

Before you start reading this unit, you should:

- Know the binomial distribution
- Have knowledge of the normal distribution
- Know to find the solutions of equations
- Find mean and variance of any distribution

Learning Objectives

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

- Understand the different terminology in sampling theory and types of sampling techniques
- Differentiate among major sampling distributions of statistics
- Determine an appropriate sample size to estimate a population mean or proportion for a given label of accuracy and with a prescribed level of confidence

INTRODUCTION

Many a times, it is of much interest to draw valid conclusions about a large group of individuals or objects. Instead of examining the entire group which may be difficult or impossible to do, we examine a small part of this entire group. This is done to draw some interfaces regarding the large group from the results of the small part. This process is called statistical inference and the process of obtaining samples is called sampling.

6.1 SOME DEFINITIONS

Population

A population consists of totality of the observations with which we are concerned.

Sample

A sample is a subset of population. Suppose we are interested in the average of most of the certain oldest reputed college in a city, the college strength can be considered as population and a small subset collected from various branches can be considered as a sample.

Parameters

The statistical measures regarding a population are called parameters. Some quantities such as mean, variance, and standard deviation of a population are referred to as parameters of population.

For example: μ is the mean of population

σ^2 is the variance of a population

Statistic

The measures (any function of random) that are with regard to a sample are called statistics. The sample statistics are quantities obtained from a sample for the purpose estimating the population parameters.

For example: $\bar{x}$ is the mean of a sample

S^2 is the variance of a sample

Sampling With and Without Replacement

Suppose we draw an object from an urn, we may replace it or may not replace it before drawing another object again. In the case where we replace, we may have a chance of getting a particular object again and again, and in the case where we do not replace, it can come up only once.

Hence, the sampling where each member of a population may be chosen more than once is called sampling with replacements, while sampling where each member cannot be chosen more than once is called sampling without replacement.

A population which has countable number of elements is said to be a finite population.

Hence, a finite population which is sampled with replacement can be theoretically considered as an infinite population.

Now let us go through different types of sampling.

6.2 TYPES OF SAMPLING

Purposive Sampling

In this type of sampling, the sample units are selected with a definite purpose in view. Hence this type suffers from the drawback of nepotism, favouritism and does not represent a sample of the population. For example, if we want to show the ECE branch of engineering students fare well in academics, only the data of good students of ECE is collected unlike the other branches, where the data of the students is collected at random.

Random Sampling

In this type of sampling, sample units are selected at random. A random sample is a sample in which each unit of population has an equal chance of being included in it.

Suppose we want to select a sample of size n from a population of size N. Then NC_n samples are possible. A process of sampling where each of the NC_n samples has an equal chance of being selected is known as random sampling. Random samples can be obtained by the use of random number tables or by throwing of a dice, draw of lottery, etc.

Simple Sampling

Simple sampling is random sampling, in which each unit of population has an equal chance of occurrence, and this probability is independent of the previous drawings. To ensure that sampling is simple it must be done with replacement if population is finite, whereas in case of infinite population, no replacement is necessary.

Stratified Sampling

This type of sampling is the best representative of the population. In this type of sampling, population is divided into a few subgroups, each of which is homogenous within itself. These homogeneous groups are termed as strata, which differ from one another. In selecting the subgroups, homogeneity within the

subgroups and heterogeneity between the subgroups are aimed at. The sample which is the aggregate of the sampled units of each of the stratum is termed as stratified sample and the technique is called stratified sampling.

6.3 ADVANTAGES OF SAMPLING

- A sample survey is cheaper than a census survey.
- The execution of the field work and the analysis of the results can be carried out easily as the magnitude of the operations involved in a sample survey is small.
- The quality of supervision, interviewing, and other related activities are better as the scale of operations involved in a sample survey is small.

Limitations of Sampling

- Sampling leads to some errors, if these errors are too large, the results of sample survey will be of limited use.
- Complete coverage of certain units is not possible.
- When the information is needed on every unit in the population such as individuals, dwelling units, and business establishments, sample survey is not of much help as it cannot provide individual information.

6.4 SAMPLING DISTRIBUTION OF A STATISTIC

If a sample of size n is drawn from a population of size N, the number of such samples that can be drawn are $NC_n = k$ (say).

For each of these samples we can complete some statistic $\hat{t} = t(x_1, x_2, \ldots x_n)$.

Consider the following table:

Sample number	Statistic			
	$\bar{x}$	s^2	…	t
1	$\bar{x}_1$	s_1^2	...	t_1
2	$\bar{x}_2$	s_2^2	…	t_2
⋮				
k	$\bar{x}_k$	s_k^2	…	t_k

The sampling distribution of the statistic is the set of values so obtained one for each sample.

The sampling distribution of the statistic t can be obtained from the above table. For such sampling distribution, we can find mean and variance.

Mean of the statistic $t = \bar{t} = \frac{1}{k}(t_1 + t_2 + \ldots t_k)$

Variance of the statistic, $t = \text{var}(t) = \frac{1}{k}[(t_1 - \bar{t})^2 + (t_2 - \bar{t})^2 + \ldots + (t_k - \bar{t})^2]$

6.5 STANDARD ERROR

In the above topic, we have seen obtaining mean and variance of a statistic of sampling distribution. We can also think of finding standard derivation of sampling distribution of that statistic.

For example, we find mean heights of such samples of a large population. It may happen that these sample means may coincide to a greater extent. We expect to see some variability in our observed means, which occurs due to sampling error just by chance. The standard deviation of the distribution of sample means measures the extent to which we expect the means from the different samples to vary, because of this chance error occurs in the sampling process.

Definition

The standard deviation of the sampling distribution of a statistic is known as standard error of the statistic.

The standard error not only indicates the size of the chance error that has been made, but also the accuracy we are likely to get if we use a sample statistic to estimate a population parameter.

A distribution of sample mean that is less spread out (that has a less standard error) is a better estimator of the population mean than the distribution of the sample mean that is widely dispersed and has a larger standard error.

The following table gives some of the statistics and their standard errors:

S. no.	Statistic	Standard error
1.	Sample mean, $\bar{x}$	$\frac{\sigma}{\sqrt{n}}$
2.	Sample variance, s^2	$\sigma^2\sqrt{\frac{2}{n}}$
3.	Sample standard deviation, s	$\sqrt{\frac{\sigma^2}{2n}}$
4.	Observed sample proportion, 'p'	$\sqrt{\frac{PQ}{n}}$
5.	Sample correlation coefficient, (r)	$\frac{(1-\rho^2)}{\sqrt{n}}$, ρ–population correlation coefficient
6.	Difference between two sample means $(\bar{x}_1 - \bar{x}_2)$	$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$
7.	Difference between two sample standard deviations, $(s_1 - s_2)$	$\sqrt{\frac{\sigma_1^2}{2n_1} + \frac{\sigma_2^2}{2n_2}}$
8.	Difference between two sample proportions $(p_1 - p_2)$	$\sqrt{\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}}$

6.6 IMPORTANCE OF STANDARD ERROR

This plays a vital role in the large sample theory and is used in testing of hypothesis. If t is any statistic then for large samples,

$$z = \frac{t - E(t)}{\sqrt{v(t)}} \sim N(0,1)$$

that is, $z = \frac{t - E(t)}{s \cdot E(t)} \sim N(0,1)$ for large samples.

[This is dealt in detail in the next unit.]

The standard error of a statistic can be reduced by increasing the sample size, but this results in corresponding increase in cost, labour, time, etc.

6.7 SAMPLING FROM NORMAL AND NON-NORMAL POPULATIONS

The sampling distribution of a mean of a sample taken from a normal population has the following:

(i) Mean equal to the population mean $\mu_{\bar{X}} = \mu$

(ii) Standard deviation that is equal to the population standard deviation divided by the square root of the sample size, that is, $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$

However, in case the population from which a sample taken is not normally distributed. Then also, the central limit theorem states that as the sample size gets large enough, the sampling distribution of the mean can be approximated by the normal distribution.

According to central limit theorem, for a large sample, the sampling distribution of the sample mean $\bar{x}$ is approximately normal, regardless of the shape of population distribution, the mean of the sampling distribution is $\mu_{\bar{X}} = \mu$ and standard deviation is $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$.

Hence, the standard error gives a measure of dispersion of the sample means around the population mean μ. If the standard error is small, then it shows small dispersion of the sample means around the population mean (tending to concentrate more closely around μ). If the standard error increases, the values taken by the sample mean are scattered away from the population mean.

6.8 FINITE POPULATION CORRECTION (*FPC*) FACTOR

When the population is infinite, the standard error is given by $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$, that is, when a sample is selected from a finite population with replacement. However, when the population is finite, then the standard case needs a correction factor to be multiplied.

$$\therefore \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

Where N is the size of the population. This term which is multiplied to $\frac{\sigma}{\sqrt{n}}$ is called *fpc* factor.

Worked Out Examples

EXAMPLE 6.1

What is the *fpc* factor when

(i) $n = 5, N = 200$

(ii) $n = 100, N = 5000$?

Solution: The *fpc* factor is given by

(i) $fpc = \sqrt{\frac{N-n}{N-1}} = \sqrt{\frac{200-5}{200-1}} = \sqrt{\frac{195}{199}}$

$fpc = 0.989$

(ii) $fpc = \sqrt{\frac{N-n}{N-1}} = \sqrt{\frac{5000-100}{5000-1}} = \sqrt{\frac{4900}{4999}}$

$fpc = 0.99$

EXAMPLE 6.2

Suppose we are interested in 15 electronic companies of the same size. All these companies are confronting a serious problem in the form of excessive turnover of their employees. It has been found that the standard deviation of distribution of annual turnover is 60 employees. Determine the standard error of the mean by taking three electronic companies, without replacement.

Solution: Given, $N = 15$, $n = 3$, $\sigma = 60$

The standard error is given by $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} = \frac{60}{\sqrt{3}} = \sqrt{\frac{15-3}{15-1}} = 34.64(0.9258)$

$= 32.07.$

Hence, the correction factor of 0.9258 has reduced the standard error from 34.64 to 32.07.

Now let us go through sampling distribution of some important statistic such as mean, variance, difference of means, etc.

6.9 SAMPLING DISTRIBUTION OF MEANS

Suppose a random variable of n observations is taken from a normal population with mean μ and variance σ^2.

Then $\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$, where x_i is the observation of the sample and has a normal distribution with mean $\mu_{\bar{X}} = \frac{\mu + \mu + \cdots + \mu}{n} = \mu$ and a variance $\sigma^2_{\bar{X}} = \frac{\sigma^2 + \sigma^2 + \cdots + \sigma^2}{n^2} = \frac{\sigma^2}{n}$.

From central limit theorem, if $\bar{X}$ is the mean of a random sample of size n taken from a population with mean μ and finite variance σ^2 than the limiting form of the distribution of $z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$ as $n \to \infty$ is the standard normal distribution.

6.10 WHEN POPULATION VARIANCE IS UNKNOWN

The standard variable $z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$ is normally distributed if the population from which samples of size n taken is normally distributed. Here we have assumed that population variance is known and $n \geq 30$. Sometimes, the population variance may be unknown. Then we can estimate the population variance by using the sample variance. Hence in this core we can seek another distribution given by

$$t = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

This follows student's *t*-distribution with $(n - 1)$ degrees of freedom whenever the population is normally distributed.

6.11 SAMPLING DISTRIBUTION OF THE DIFFERENCE BETWEEN TWO MEANS

Suppose that we have two populations with means μ_1 and μ_2 and variances σ_1^2 and σ_2^2. Let the statistic $\bar{X}_1$ represented the mean of the random sample of size n_1 selected from the first population. Let the statistic $\bar{X}_2$ represent the mean of a random sample selected from the second population which is independent of the sample of the first population. The variables $\bar{X}_1$ and $\bar{X}_2$ are both approximately normally distributed with means μ_1 and μ_2 and variances $\frac{\sigma^2}{n_1}$ and $\frac{\sigma^2}{n_2}$, respectively.

Then $\bar{X}_1 - \bar{X}_2$ is approximately normally distributed with mean and variance given by

$$\mu_{\bar{X}_1 - \bar{X}_2} = \mu_1 - \mu_2 \text{ and } \sigma^2_{\bar{X}_1 - \bar{X}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

and hence, $z = \dfrac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\left(\dfrac{\sigma_1^2}{n_1}\right) + \left(\dfrac{\sigma_2^2}{n_2}\right)}}$ is approximately normal distribution.

6.12 SAMPLING DISTRIBUTION OF VARIANCE

If a random sample of size n is drawn from a normal population with mean μ and variance σ^2 and the sample variance is computed, we obtain the value of the statistic s^2.

$$s^2 = \frac{(X_1 - \bar{X})^2 - (X_2 - \bar{X})^2 + \cdots + (X_n - \bar{X})^2}{n}$$

If s^2 is the variance of a random sample of size n taken from a normal population having the variance σ^2, then the statistic

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2} = \sum_{n=1}^{n} \frac{(X_i - \bar{X})^2}{\sigma^2}$$

has a Chi-squared distribution with $n - 1 = \gamma$ degrees of freedom.

Degrees of Freedom

The quantity $(n - 1)$ is called degrees of freedom associated with the variance estimated. It depicts the number of independent pieces of information available for computing variability.

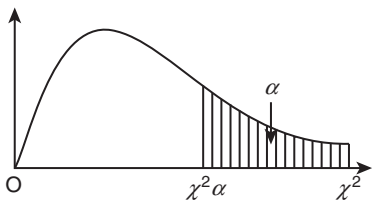
Thus, there are $(n - 1)$ degrees of freedom for computing a sample variance rather than n degrees of freedom.

6.13 THE CHI-SQUARE DISTRIBUTION

The Chi-square values are calculated from each sample by the formula

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

The probability that a random sample produces a χ^2 value greater than some specified value is equal to the area under the



curve to the right of this value. Here, χ^2_α represents the value of χ^2 above which we find area α, given in the shaded portion of the following figure.

EXAMPLE 6.3

A population consists of five elements 2, 3, 6, 8, and 11. Consider all possible samples of size 2 that can be drawn without replacement from this population. Find:

(i) Mean of the population
(ii) The standard dev iation of the population
(iii) The mean of the sampling distribution of means and
(iv) Standard deviation of sampling distribution of means
(v) Repeat the above problem from (i) to (iv) by considering the samples with replacement

Solution:

Without replacement:

This is a finite population. Hence, the total number of samples without replacement is $NC_n = 5C_2 = 10$, since population size = 5 and sample size = 2.

There are 10 samples of size 2. The following are the 10 samples:

(2, 3) (2, 6) (2, 8) (2, 10)
(3, 6) (3, 8) (3, 11)
(6, 8) (6, 11)
(8, 11)

Here (2, 3) is considered the same as (3, 2). The corresponding sample means are

2.5 4 5 6
4.5 5.5 7
7 8.5
9.5

(i) Mean of the population

$$\mu = \frac{3+3+6+8+11}{5} = \frac{30}{5} = 6$$

(ii) Standard deviation of the population

Variance of the population, $\sigma^2 = \dfrac{\sum (x_i - \bar{x})^2}{n}$

$$= \frac{1}{5}[(2-6)^2 + (3-6)^2 + (6-6)^2 + (8-6)^2 + (11-6)^2]$$

$$\sigma^2 = 10.8$$

$\therefore$ Standard deviation of the population = $\sigma = 3.29$

(iii) The sampling distribution of means can be given as follows:

Sampling distribution of means:

x_i	2.5	4	4.5	5	5.5	6.5	7	8.5	9.5
f_i	1	1	1	1	1	1	2	1	1
$f_i x_i$	2.5	4	4.5	5	5.5	6.5	14	8.5	9.5

Mean of sampling distribution of means:

$$\mu_{\bar{X}} = \frac{\Sigma f_i x_i}{\Sigma f_i}$$
$$= \frac{60}{10}$$
$$= 6$$

Variance of the sampling distribution of means:

$$\sigma^2_{\bar{X}} = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right) = \frac{10.8}{2}\left(\frac{5-2}{5-1}\right)$$
$$= 4.05$$

This is the variance of sampling without replacement. Standard deviation of sampling distribution of means $\sigma_{\bar{X}} = 2.0124$

Sampling with replacement:

This is infinite population. The total number of samples with replacement = $N^n = 5^2 = 25$

Sample size 2 are as follows:

(2, 2)	(2, 3)	(2, 6)	(2, 8)	(2, 11)
(3, 2)	(3, 3)	(3, 6)	(3, 8)	(3, 11)
(6, 2)	(6, 3)	(6, 6)	(6, 8)	(6, 11)
(8, 2)	(8, 3)	(8, 6)	(8, 8)	(8, 11)
(1, 2)	(11, 3)	(11, 6)	(11, 8)	(11, 11)

The sample means of these samples are as follows:

2	2.5	4	5	6.5
2.5	3	4.5	5.5	7
4	4.5	6	7	8.5
5	5.5	7	8	9.5
6.5	7	8.5	9.5	11

The sample distribution of means is as follows:

x_i	2	2.5	3	4	4.5	5	5.5	6	6.5	7	8	8.5	9.5	11
f_i	1	2	1	2	2	2	2	1	2	4	1	2	2	1
$x_i f_i$	2	5	3	8	9	10	11	6	13	28	8	17	19	11

Mean of the sampling distribution of means is given by,

$$\mu_{\bar{x}} = \frac{\Sigma f_i x_i}{\Sigma f_i}$$
$$= \frac{150}{25}$$
$$= 6$$

Variance of the sampling distribution of means is given by,

$$= \frac{\sigma^2}{n}$$
$$= \frac{10.8}{5}$$
$$= 2.16$$

Hence, the standard deviation of sampling distribution of means = standard error of means = 1.469.

EXAMPLE 6.4

The amount of time that a drive-through bank teller spends on a customer is a random variable with a mean $\mu = 3.2$ minutes and a standard deviation $\sigma = 1.6$ minutes. If a random sample of 64 customers is observed, find the probability that their mean time at the teller's counter is

(i) At most 2.7 minutes

(ii) More than 3.5 minutes

(iii) At least 3.2 minutes but less than 3.4 minutes.

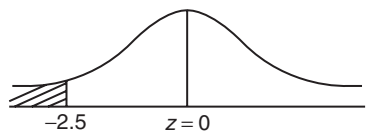
Solution: From data given, $\mu = 3.2$, standard deviation $\sigma = 1.6$ minutes, size of the random sample, $n = 64$.

Let x denotes the time at the teller's counter,

(i) Probability that mean time at the teller's counter is at most 2.7 minutes = $P(\bar{x} < 2.7)$

$$z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \text{ when } \bar{X} = 2.7, \quad z = \frac{2.7 - 3.2}{\frac{1.6}{\sqrt{64}}} = \frac{-0.5}{0.2} = -2.5$$

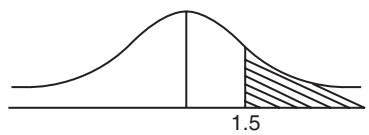
$$P(\bar{X} < 2.7) = P(z < -2.5)$$
$$= P(z > 2.5)$$
$$= 0.0062$$



(ii) Probability that mean time is more than 3.5 minutes $P(\bar{x} > 3.5)$

$$z = \frac{\bar{x} - \mu}{\sigma\sqrt{n}} = \frac{3.5 - 3.2}{\frac{1.6}{\sqrt{64}}} = \frac{0.3}{0.2} = 1.5$$

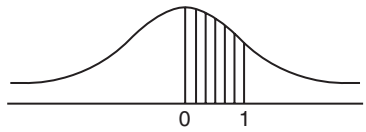
$$P(\bar{x} > 3.5) = P(z > 1.5)$$
$$= 1 - 0.9332$$
$$= 0.0668$$



(iii) Probability that mean time is at least 3.2 minutes but less than 3.4 minutes $= P(3.2 < \bar{X} < 3.4)$

$$\bar{X} = 3.2,\ z = \frac{3.2 - 3.2}{\frac{1.6}{\sqrt{64}}} = 0, \quad \bar{X} = 3.4,\ z = \frac{3.4 - 3.2}{\frac{1.6}{\sqrt{64}}} = \frac{0.2}{0.2} = 1$$

$$\begin{aligned} P(3.2 < \bar{X} < 3.4) &= P(0 < z < 1) \\ &= 0.8413 - 0.5 \\ &= 0.3413 \end{aligned}$$



EXAMPLE 6.5

For a Chi-squared distribution, find the following:

(i) $\chi^2_{0.005}$ when $\gamma = 5$

(ii) $\chi^2_{0.05}$ when $\gamma = 19$

(iii) $\chi^2_{0.01}$ when $\gamma = 12$

Solution:

(i) $\chi^2_{0.005}$ when $\gamma = 5$ degrees of freedom is a $\chi^2_{0.005} = 16.750$

(ii) $\chi^2_{0.05}$ when $\gamma = 19$ degrees of freedom is $\chi^2_{0.05} = 30.144$

(iii) $\chi^2_{0.01}$ when $\gamma = 12$ degrees of freedom is $\chi^2_{0.01} = 26.216$

EXAMPLE 6.6

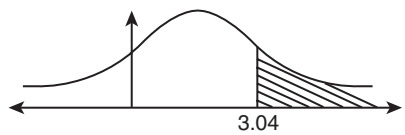
A random sample of size 25 is taken from a normal population having a mean of 80 and a standard deviation of 5. A second random sample of size 36 is taken from a different normal population having mean of 75 and standard deviation of 3. Find the probability that the sample mean computed from the 36 measurements by at least 3.4 but less than 5.9.

Solution: Probability that the sample mean computed from 25 measurements will exceed sample mean computed from 36 measurements by at least 3.4 but less than 5.9 $= P(3.4 < \bar{X}_A - \bar{X}_B < 5.9)$

$$z = \frac{\bar{X}_A - \bar{X}_B - (\mu_A - \mu_B)}{\sqrt{\frac{\sigma_{A^2}}{n_1} + \frac{\sigma_{B^2}}{n_2}}}$$

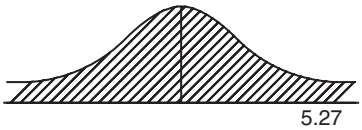
$$P(\bar{X}_A - \bar{X}_B > 3.4) = \frac{3.4}{\sqrt{\frac{5^2}{25} + \frac{3^2}{36}}} = \frac{3.4}{1.118}$$

$$\begin{aligned} P(\bar{x}_A - \bar{x}_B > 3.4) &= P(z > 3.041) \\ &= 1 - 0.9988 \\ &= 0.0012 \end{aligned}$$



$$P(\bar{X}_A - \bar{X}_B < 5.9) = \frac{5.9}{1.118} = 5.277 = P(z < 5.277)$$

$$= 0.9998$$



EXAMPLE 6.7

Ball bearings of a given brand weigh 0.50 oz with a standard deviation of 0.02 oz. What is the probability that two lots of 1,000 ball bearings each, will differ in weight by more than 2 oz?

Solution: Let $\bar{X}_1$ and $\bar{X}_2$ be the mean weights of ball bearings in the two lots. Then

$$\mu_{\bar{X}_1 - \bar{X}_2} = \mu_{\bar{X}_1} - \mu_{\bar{X}_2} = 0.50 - 0.50 = 0$$

$$\sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\frac{(0.02)^2}{1000} + \frac{(0.02)^2}{1000}} = 0.000895$$

The standard normal variable is

$$z = \frac{\bar{X}_1 - \bar{X}_2 - \mu_{\bar{X}_1 - \bar{X}_2}}{\sigma_{\bar{X}_1 - \bar{X}_2}} = \frac{\bar{X}_1 - \bar{X}_2 - 0}{0.000895}$$

Given that two lots differ by $20z = \frac{2}{1000} = 0.002$ oz in the means (i.e.,) $\bar{X}_1 - \bar{X}_2 \geq 0.002$ or $\bar{X}_1 - \bar{X}_2 \leq -0.02$

$$z \geq \frac{0.002}{0.000895} = 2.23 \text{ or } z \leq \frac{-0.002}{0.000895} = -2.23$$

$\therefore P(z \geq 2.23 \text{ or } z \leq -22.3) = P(z \geq 2.23) + P(z \leq -22.3) = 0.0258$

EXAMPLE 6.8

The mean score of students on an aptitude test is 72 points with a standard deviation of 8 points and another test with mean 86 and standard deviation of 11. What is the probability that two groups of students consisting of 28 and 36 students respectively, will differ in their mean scores by

(i) 3 or more points,

(ii) 6 or more points,

(iii) Between 2 and 5 points?

Solution:

(i) Probability that mean scores of students of an aptitude test are more than 3

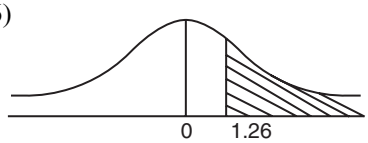
$$= P(\bar{X}_A - \bar{X}_B > 3)$$

$$= P\left(z > \frac{\bar{X}_A - \bar{X}_B}{\sqrt{\frac{J_A^2}{n_1} + \frac{J_B^2}{n_2}}} \right)$$

$$= P\left(z > \frac{3}{\sqrt{\frac{8^2}{28} + \frac{11^2}{36}}}\right) = P\left(z > \frac{3}{\sqrt{2.28 + 3.361}}\right)$$

$$= P\left(z > \frac{3}{2.376}\right) = P(z > 1.26)$$

$$= 1 - 0.8962 = 0.1038$$



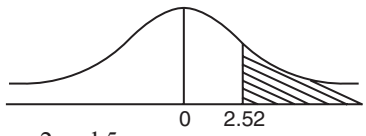
(ii) Probability that mean scores of students in the test are more than 6

$$P(\bar{X}_A - \bar{X}_B > 6) = P\left(z > \frac{6}{2.376}\right)$$

$$= P(z > 2.52)$$

$$= 1 - 0.9941$$

$$= 0.0059$$



(iii) Probability that mean scores of students in the test are between 2 and 5

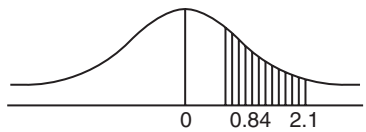
$$= P(2 < \bar{X}_A - \bar{X}_B < 5)$$

$$= P\left(\frac{2}{2.376} < z < \frac{5}{2.376}\right)$$

$$= P(0.841 < z < 2.104)$$

$$= 0.9821 - 0.7995$$

$$= 0.1826$$



EXAMPLE 6.9

Find the probability that a random sample of 25 observations from a normal population with variance $\sigma^2 = 6$ will have a variance s^2

(i) Greater than 9.1 and

(ii) Between 3.462 and 10.745.

Solution: Given that n = sample size = 25

Population variance, $a^2 = 6$

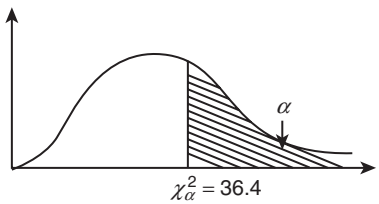
(i) Sample variance, $s^2 = 9.1$

Hence, the probability that sample variance will be greater than 9.1 is $P(\chi^2 > \chi^2_\alpha)$

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2} = \frac{(25-1)(9.1)}{6}$$

$$= \frac{24(9.1)}{6} = 36.4$$

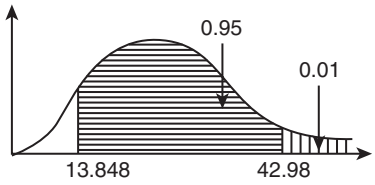
$$\therefore P(\chi^2 > 36.4) = ?$$



From tables it is clear that for $\gamma = n - 1 = 2n$ degree of freedom, $P(\chi^2 > 36.4) = 0.05$

(ii) When $s^2 = 3.462$

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2} = \frac{24(3.462)}{6} = 13.848$$



When $s^2 = 10.745$

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2} = \frac{24(10.745)}{6} = 42.98$$

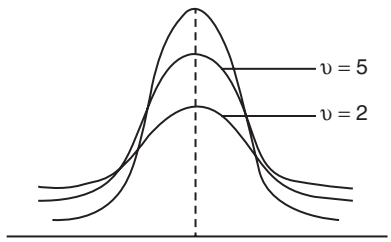
$$\therefore P(13.848 < \chi^2 < 42.98) = P(\chi^2 > 13.848) - P(\chi^2 < 42.98)$$

$$0.95 - 0.01 = 0.94$$

6.14 THE STUDENT'S *t*-DISTRIBUTION

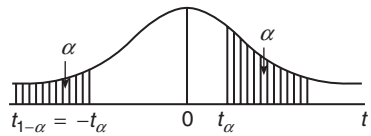
Let $X_1, X_2, \ldots X_n$ be an independent random variable that are all normal with mean μ and standard deviation σ. Let $\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$ and $s^2 = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n-1}$

Then the random variable $t = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}}$ has a *t*-distribution with $\gamma = n - 1$ degrees of freedom. The *t*-distribution is almost similar to *z* distribution. This is bell-shaped and symmetric about mean of zero.



The *t*-distribution for $\upsilon = 2, 5, \ldots$

Only difference is that the variance of *t* depends on the sample size *n* and is always greater than 1. The t_α represents the *t*-value above which we find an area equal to α. We can observe that



$$t_{1-\alpha} = -t_\alpha$$

Hence,

$$t_{0.95} = -t_{0.05}$$
$$t_{0.99} = -t_{0.01}$$

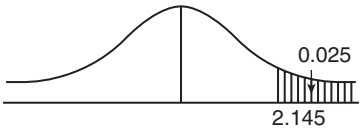
EXAMPLE 6.10

(i) Find $t_{0.025}$ when $\gamma = 14$

(ii) Find $t_{-0.10}$ when $\gamma = 10$

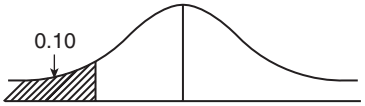
Solution:

(i)



From tables $t_{0.025} = 2.145$ when $\gamma = 14$ degrees of freedom

(ii)



$-t_{0.10} = t_{0.90}$,

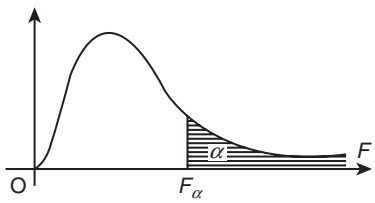
$-t_{0.10} = 1.372$, for $\gamma = 10$ degrees of freedom

6.15 *F*-DISTRIBUTION

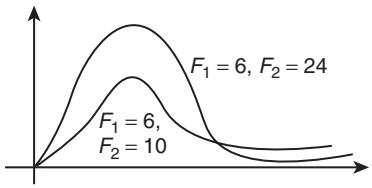
Let x_i, $i = 1, 2, \ldots n_1$, and y_j, $j = 1, 2, \ldots n_2$ be two samples which are independent then the random variable $F = \frac{s_x^2}{s_y^2}$ follows *F*-distribution with $(\gamma_1, \gamma_2) = (n_1 - 1, n_2 - 1)$ degrees of freedom, where

$$s_x^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n}(x_i - \bar{x})^2,\ s_Y^2 = \frac{1}{n_2 - 1}\sum_{j=1}^{n_2}(y_j - \bar{y})^2$$

The critical values of *F*-distribution are calculated as follows. F_α is the value of *F*-statistic above which there is an area α.



As an example, F_α is shown for $\gamma_1 = 6$ and $\gamma_2 = 10$ degrees of freedom and $\gamma_1 = 6$ and $\gamma_2 = 24$ degrees of freedom leaving. The *f*-values are tabulated for $\alpha = 0.05$ and $\alpha = 0.01$ for various combination of degrees of freedom leaving an area of 0.05 to the right. Hence, $f_{0.05} = 3.22$ for $\gamma_1 = 15$ and $\gamma_2 = 8$ degrees of freedom.



In addition, $F_\alpha(\gamma_1, \gamma_2)$ for F_α with $\gamma_1 = n_1 - 1$ and $\gamma_2 = n_{2-1}$ degrees of freedom, we have

$$F_{1-\alpha}(\gamma_1, \gamma_2) = \frac{1}{F_2(\gamma_2, \gamma_1)}$$

$$\text{Hence, } F_{0.95}(6, 10) = \frac{1}{F_{0.05}(10, 6)} = \frac{1}{4.06}$$

$$= 0.246$$

Worked Out Examples

EXAMPLE 6.11

Find F for the following:

(i) $\gamma = 7$, $\gamma_2 = 15$ for 5%, 1% level of significance

(ii) $\gamma_1 = 10$ $\gamma_2 = 15$ for $\alpha = 1\%$, 5%

(iii) $\gamma_1 = 15$, $\gamma_2 = 7$ for $\alpha = 0.95$

Solution: For $\alpha = 5\%$ level of significance

(i) $F_{0.05}(7-1, 15-1) = F_{0.05}(6, 14) = 2.85$

$F_{0.01}(7-1, 15-1) = F_{0.01}(6, 14) = 4.46$

(ii) $F_{0.05}(10-1, 15-1) = F_{0.05}(9, 14) = 2.65$

$F_{0.01}(10-1, 15-1) = F_{0.01}(9, 14) = 4.03$

(iii) $F_{0.95}(15-1, 7-1) = F_{0.95}(14, 6)$

$$= \frac{1}{F_{0.05}(6,14)} = \frac{1}{2.85}$$

$$= 0.3508$$

Work Book Exercises

1. A normal variate has mean 0.5 and standard deviation 2.5. Find the probability that
   (i) Mean of a random sample of size 16 from the population is more than 1.6
   (ii) Mean of the random sample of size 90 from the population is negative

2. The guaranteed average life of certain type of electric bulb is 1,000 hours with a standard deviation of 125 hours. To ensure 90% of bulbs do not fall short of the guaranteed average by more than 2.5%, find the minimum size of the sample which satisfies this condition.

   [**Ans.:** $n = 41$]

3. If the distribution of the weight of all men travelling by air between Dallas and Elapse has a mean of 163 pounds and a standard deviation of 18 pounds. What is the probability that the combined gross weight of 36 men travelling on a plane between these two cities is more than 6000 pounds?

4. A population consists of 4 elements 1, 5, 6, and 9. Consider all possible samples of size 3, that can be drawn without replacement, from the population. Find the following:
   (i) Mean of the population
   (ii) Standard deviation of the population
   (iii) Mean of the sampling distribution of means
   (iv) Standard deviation of the sampling distribution of means.

5. The average life of a bread-making machine is seven years with a standard deviation of one year. Assuming that the lives of these machines follow approximately a normal distribution. Find:
   (i) The probability that the mean life of a random sample of 16 such machines falls between 6.4 and 7.2 years.
   (ii) The value of $\bar{x}$ to the right of which 15% of the means computed from random samples of first problem fall.

6. For a Chi-squared distribution, find the following:
   (i) $\chi^2_{0.01}$ when $\gamma = 12$
   (ii) $\chi^2_{0.005}$ when $\gamma = 5$
   (iii) $\chi^2_{0.05}$ when $\gamma = 24$

7. Find the probability that random sample of 25 observations from a normal population with variance $\alpha^2 = 6$ will have variance
   (i) Greater than 7.1
   (ii) Between 8.462 and 10.740

8. For an *F*-distribution, find:
   (i) $F_{0.99}$ with $\gamma_1 = 28, \gamma_2 = 12$
   (ii) $F_{0.01}$ with $\gamma_1 = 24, \gamma_2 = 19$
   (iii) $F_{0.0,5}$ with $\gamma_1 = 19, \gamma_2 = 24$

9. For a *t*-distribution, find
   (i) $t_{0.025}$ when $\gamma = 15$
   (ii) $t_{0.995}$ when $\gamma = 4$
   (iii) $t_{0.05}$ when $\gamma = 22$

10. Find *k* such that
   (i) $P(-k < t < k) = 0.90$
   (ii) $P(k < t < 2.807) = 0.095$
   (iii) $P(-2.069 < t < k) = 0.965$

11. A research worker wishes to estimate mean of a population by using sufficiently large sample. The probability is 95% that sample mean will not differ from the true mean by more than 25% of the standard deviation. How large a sample should be taken?

Caution: If the sample size is small $n < 30$, the *t*-distribution gives good inferences on μ given by $t = \dfrac{\bar{x} - \mu}{\dfrac{s}{\sqrt{n}}}$. Hence, when the central limit theorem is used, when the population standard deviation is unknown, we use *t*-distribution with an estimate for σ.

DEFINITIONS AT A GLANCE

Population: Totality of observations with which we are concerned
Sample: Subset of population
Statistic: Measures of sample
Parameter: Measures of a population
Standard Error: Standard deviation of sampling distribution of a statistic

FORMULAE AT A GLANCE

- When the population is finite, then the standard error needs a correction factor given by

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$

- From central limit theorem, the standard normal distribution is $z = \dfrac{\bar{X} - \mu}{\dfrac{\sigma}{\sqrt{n}}}$.
- The sampling distribution of difference between two means is $z = \dfrac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$.
- The sampling distribution of variance is $\chi^2 = \dfrac{(n-1)s^2}{\sigma^2}$ follows χ^2 distribution with $\gamma = n - 1$ degrees of freedom.
- When the population variance is unknown, then $t = \dfrac{\bar{X} - \mu}{\dfrac{s}{\sqrt{n}}}$.

The difference of population variances is $F = \dfrac{s_X^2}{s_Y^2}$ follows F-distribution with $(\gamma_1, \gamma_2) - (n_1 - 1, n_2 - 1)$ degrees of freedom and $s_X^2 = \dfrac{1}{n_1 - 1}\Sigma(x_i - \bar{x})^2$, $s_Y^2 = \dfrac{1}{n_2 - 1}\Sigma(y_j - \bar{y})^2$.

OBJECTIVE TYPE QUESTIONS

1. The standard error of sample standard deviation is

   (a) $\sqrt{\dfrac{\sigma^2}{n}}$ (b) $\sqrt{\dfrac{\sigma^2}{2n}}$

   (c) $\dfrac{\sigma}{n}$ (d) none

2. The standard error of a statistic t is $\dfrac{\sigma}{\sqrt{n}}$, then t is

   (a) sample variance (b) sample s. d

   (c) sample mean (d) none

3. The *fpc* factor is given by

   (a) $\sqrt{\frac{\sigma}{n}}$ (b) $\sqrt{\frac{N-n}{N-1}}$

   (c) $\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$ (d) none

4. If $n = 5$, $N = 100$ and $\sigma = 1.849$ then *fpc* is

   (a) 0.81 (b) 0.979
   (c) 0.61 (d) none

5. When the population is finite, the standard error of $\bar{x}$ is

   (a) $\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$ (b) $\frac{\sigma}{\sqrt{n}}$

   (c) $\sqrt{\frac{N-1}{N-n}}$ d) none

6. According to central limit theorem, $z =$

   (a) $\bar{x} - n$ (b) $\frac{\bar{x}-\mu}{\sigma}$

   (c) $\frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}}$ (d) none

7. The sampling distribution of variance is

   (a) χ^2 variate with (n – 1) degrees of freedom
   (b) χ^2 variate with n degrees of freedom
   (c) z variate
   (d) none

8. The sampling distribution of difference of two means, $z =$

   (a) $\frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{\sigma_1^2 + \sigma_2^2}}$ (b) $\frac{(\bar{X}_1 - \bar{X}_2)(\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$

   (c) $\frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{2n_1} + \frac{\sigma_2^2}{2n_2}}}$ (d) $\frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{2\sigma_1^2}{n_1} + \frac{2\sigma_2}{n_2}}}$

9. The χ^2-distribution for population variance is

   (a) $\chi^2 = \frac{(n-1)}{\sigma^2}$ (b) $\chi^2 = \frac{s^2}{\sigma^2}$

   (c) $\chi^2 = \frac{(n-1)s^2}{\sigma^2}$ (d) none

10. For a t-distribution, $-t_{0.025}$ for $\gamma = 14$ degrees of freedom given $t_{0.975} = -2.145$
    (a) –2.145 (b) –3.841
    (c) –1.96 (d) none

11. The t-distribution for a small sample mean is
    (a) $t = \dfrac{\bar{x} - \mu}{n}$ (b) $t = \dfrac{\bar{x} - \mu}{\dfrac{s}{\sqrt{n}}}$
    (c) $t = \dfrac{\bar{x} - \mu}{\dfrac{1}{\sqrt{n}}}$ (d) none

12. The total number of samples without replacement for $N = 5$, $n = 2$ is
    (a) 10 (b) 5
    (c) 25 (d) none

13. The total number of samples with replacement for $N = 5$, $n = 2$ is
    (a) 25 (b) 10
    (c) 5 (d) none

ANSWERS

1. (b) 2. (c) 3. (c) 4. (a) 5. (b) 6. (c) 7. (a) 8. (b)
9. (c) 10. (a) 11. (b) 12. (a) 13. (a)

7 Testing of Hypothesis (Large Samples)

Prerequisites

Before you start reading this unit, you should:

- Be familiar with sampling theory
- Have knowledge of sampling distributions

Learning Objectives

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

- Understand the concepts of hypothesis null and alternative hypothesis
- Understand the procedure for testing of hypothesis
- Distinguish between Type I and Type II errors in hypothesis testing
- Use an appropriate test for testing situations involving means, proportions, and the differences between means and proportions.

INTRODUCTION

In Unit 6, we have seen the sampling distributions and finding the mean and variance for such distributions. In this unit, we shall draw inferences for deciding about the characteristics of the populations on the basis of the sample study. First let us define the term statistical hypothesis.

7.1 STATISTICAL HYPOTHESIS

A statistical hypothesis is an assertion or conjecture or a statement concerning one or more population.

The truth or falsity of a statistical hypothesis is known only when we include the entire population. This is impossible in most situations. Hence, we consider a random sample from the population and use the data in this sample to provide evidence that supports or does not support the hypothesis.

7.2 TESTS OF SIGNIFICANCE

The study of tests of significance is very important in sampling theory which helps us to decide based on the sample results,

(i) If the difference between the observed sample statistics and the hypothetical parameter value, or

(ii) If the difference between two independent sample statistics,

is significant or it can be attributed to chance or fluctuations in sampling.

For large n, almost all the distributions such as Binomial, Poisson, Negative Binomial, and Hypergeometric can be approximated very closely by a normal probability curve. Hence, we use normal test for large samples.

Some of the terms are required, which are defined as follows:

7.3 SOME IMPORTANT DEFINITIONS

Null Hypothesis

For applying the tests of significance, we first set up a hypothesis which refers to a definite statement. We wish to test about the population parameter. Such a hypothesis of no difference is called null hypothesis. It is usually denoted by H_0.

Alternative Hypothesis

Any statement which is complementary to the null hypothesis is called alternative hypothesis. It is usually denoted by H_1.

For example, in case of a single statistic, H_0 will be that the sample statistic does not differ from the hypothetical parameter value and in case of two statistics, H_0 will be the sample statistics that do not differ significantly.

If we want to test the null hypothesis, that the population has a specified mean μ_0, that is, $H_0: \mu = \mu_0$ then the alternative hypothesis could be

(i) $H_1: \mu \neq \mu_0$ (i.e., $\mu > \mu_0$ or $\mu < \mu_0$)

(ii) $H_1: \mu > \mu_0$

(iii) $H_1: \mu < \mu_0$

The alternative hypothesis in (i) is known as two-tailed alternative and that in (ii) is known as right-tailed alternative and that in (iii) is known as left-tailed alternative, respectively.

Test Statistic

We now compute test statistic, which is based on an appropriate probability distribution. It is used to test whether the null hypothesis set up should be accepted or rejected. For this purpose, we use Z distribution under normal curve for large samples where the sample size is equal to or greater than 30 (i.e., $n \geq 30$).

Errors in Sampling

The main objective of sampling theory is to draw valid inference about the population parameters based on sample results. In this process, we decide to accept or reject the lot after examining a sample from it. Hence, we may commit the following two types of errors here:

Type I error: Reject null hypothesis H_0 when it is true.

Type II error: Accept null hypothesis H_0 when it is not true or accept H_0 when alternative hypothesis H_1 is true.

Actual	Decision	
	Accept H_0	Reject H_0
H_0 is true	Correct Decision (No error) Probability = $1 - \alpha$	Wrong Decision (Type I error) Probability = α
(H_1 is true) or H_0 is false	Wrong Decision (Type II Error) Probability = β	Correct Decision (No error) Probability = $1 - \beta$

In order to decide whether to accept or reject a null hypothesis, we wish to minimize the probability of making Type I error.

$\therefore P\{H_0 \text{ when it is true} = P\left\{\frac{H_0}{H_0}\right\} = \alpha$

$\therefore$ Probability of making a correct decision = $1 - \alpha$.

In addition, $P\{\text{accept } H_0 \text{ when it is wrong}\} = P\left\{\text{accept } \frac{H_0}{H_1}\right\} = \beta$

Probability of correct decision = $1 - \beta$

The sizes of Type I and Type II errors are given by α and β, respectively.

Caution:

- Type I error is rejecting a lot when it is good and Type II error is accepting a lot when it is bad.
- α and β are called as producer's risk and consumer's risk, respectively.

Level of Significance

The next important task is to fix level of significance, which is usually done in advance. The maximum probability of making Type I error is referred to as level of significance.

In practise, the commonly used levels of significance are 5%, (0.05), 1% (0.01), and 10% (0.1).

Critical Region

The region of the standard normal curve corresponding to a predetermined level of significance that is fixed for knowing the probability of making Type I error of rejecting the hypothesis which is true is known as critical region. In other words, a region in the sample space which amounts to rejection of H_0 is called critical region. This region is also called rejection region. The region of the normal other than rejection region is acceptance region.

One-tailed and Two-tailed Tests

A test of a statistical hypothesis where the alternative hypothesis is one-tailed (right-tailed or left-tailed) is called one-tailed test.

For example, in testing, if the population mean is μ_0, that is, H_0: $\mu = \mu_0$ against the alternative hypothesis H_1: $\mu > \mu_0$ (right-tailed) or H_1: $\mu < \mu_0$ (left-tailed) is a single tailed test.

A test of a statistical hypothesis where the alternative hypothesis is two-tailed is called a two-tailed test.

For example, in testing if a population has a mean μ_0, that is, H_0: $\mu = \mu_0$ against the alternative hypothesis H_1: $\mu \neq \mu_0$ ($\mu > \mu_0$ and $\mu < \mu_0$) is a two-tailed test.

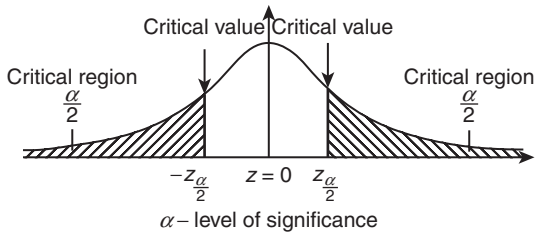
Critical Value

The value of the test statistic that separates the critical region and acceptance region is called critical value or significant value.

This value depends upon the following two factors:

(i) The level of significance used

(ii) The alternative hypothesis whether it is single tailed or two-tailed test



The critical region and critical values in two-tailed test are given as follows:

The critical value of the test statistic at α level of significance is given by $Z_{\frac{\alpha}{2}}$, that is, $Z_{\frac{\alpha}{2}}$ is the value so that the total area of the critical region on both the tails is α, that is, we have

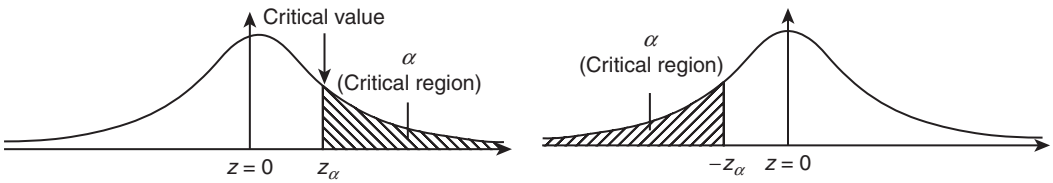
$P(Z > Z_{\frac{\alpha}{2}}) + P(Z < -Z_{\frac{\alpha}{2}}) = \alpha$ [since normal curve is symmetric]

$$2P(Z > Z_{\frac{\alpha}{2}}) = \alpha$$

$$\therefore P(Z > Z_{\frac{\alpha}{2}}) = \frac{\alpha}{2}$$

(i.e.,) the area of each tail is $\frac{\alpha}{2}$ and Z_α is the value of Z such that the area to the right of Z_α is $\frac{\alpha}{2}$ as shown in the following diagram.

Similarly, the critical region and critical values in case of one-tailed test are as follows:



For a right-tailed test, the area to the right of Z_α is α, the critical region (i.e.,)

$$P(Z > Z_\alpha) = \alpha$$

For a left-tailed test, the area to the left of z_α is α, the critical region (i.e.,)

$$P(Z < -Z_\alpha) = \alpha$$

Hence, from the above theory it is clear that, the critical value of Z for a one-tailed test (left or right-tailed) at α level of significance is same as the critical value of Z for a two-tailed test at level of significance 2α.

The table of critical values for different levels of significance and for one- and two-tailed tests is as follows:

Critical value Z_α	Level of significance (α)			
	1%	2%	5%	10%
Two-tailed test	$\lvert Z_\alpha \rvert = 2.58$	$\lvert Z_\alpha \rvert = 2.33$	$\lvert Z_\alpha \rvert = 1.96$	$\lvert Z_\alpha \rvert = 1.645$
Right-tailed test	$Z_\alpha = 2.33$	$Z_\alpha = 2.05$	$Z_\alpha = 1.645$	$Z_\alpha = 1.28$
Left-tailed test	$Z_\alpha = -2.33$	$Z_\alpha = -2.05$	$Z_\alpha = -1.645$	$Z_\alpha = -1.28$

7.4 STEPS INVOLVED IN TESTING OF HYPOTHESIS

The following are the steps involved in testing of hypothesis:

1. **Step-1:** Null hypothesis—set up the null hypothesis H_0.
2. **Step-2:** Alternative hypothesis—set up the alternative hypothesis H_1. This helps us to decide whether we need to use a one-tailed (right-tailed or left-tailed) or two-tailed test.

3. **Step-3:** Level of significance—this is fixed an advance. An appropriate level of significance is to be chosen.
4. **Step-4:** Test statistic—we compute the test statistic or test criterion.
5. Under H_0, $Z = \dfrac{t - E(t)}{\sqrt{\text{var}(t)}}$ follows normal distribution $N(0, 1)$, where t is the test statistic.
6. **Step-5:** Conclusion—after computing the calculated value of Z, we now find the tabulated value of Z, that is, critical value Z_α for α level of significance.

We compare calculated value of Z with tabulated value of Z, Z_α.

If $|Z| < Z_\alpha$, that is, the calculated value of Z is less than the tabulated value of Z, then we accept null hypothesis H_0, that is, we say that the difference $t - E(t)$ is first due to fluctuations of sampling.

If $|Z| > Z_\alpha$, that is, if the calculated value of Z is greater than the tabulated value of Z then we reject null hypothesis H_0 and hence accept null hypothesis H_0, that is, the difference is significant.

7.5 TESTS OF SIGNIFICANCE

Type I: Testing of Significance of a Single Mean (Large Sample)

We know that if the test statistics is $\bar{x}$, then

$$E(\bar{x}) = \mu$$

and

$$V(\bar{x}) = \frac{\sigma^2}{n}$$

or

Standard error of $\bar{x} = \dfrac{\sigma}{\sqrt{n}}$

$\therefore$ Under the null hypothesis, H_0, the standard normal variate corresponding to $\bar{x}$ is

$$z = \frac{\bar{x} - \mu}{\dfrac{\sigma}{\sqrt{n}}} \sim N(0, 1) .$$

Worked Out Examples

EXAMPLE 7.1

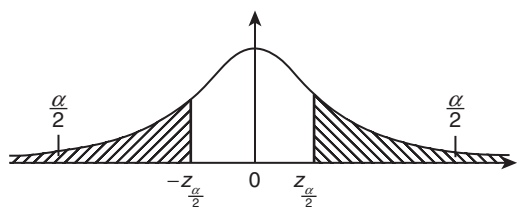
A sample of size 400 was drawn and the sample mean was found to be 99. Test whether this sample could have come from a normal population with mean 100 and variance 64 at 5% level of significance.

Solution:

(i) **Null hypothesis:** The sample has been taken from a normal population with mean $\mu = 100$, that is, H_0: $\mu = 100$.

(ii) **Alternative hypothesis:** H_1: $\mu \neq 100$ (two-tailed test).

(iii) **Level of significance:** $\alpha = 5\%$ level. Hence the critical value of Z at 0.05 level of significance is $Z_{\frac{\alpha}{2}} = \pm 1.96$



(iv) **Test statistic:** Under H_0, $Z = \dfrac{\bar{x} - \mu}{\dfrac{\sigma}{\sqrt{n}}}$

From the given data, $\bar{x} = 99$, $n = 400$, $\sigma^2 = 64$

$$Z = \frac{100 - 99}{\dfrac{8}{\sqrt{400}}} = \frac{1}{0.4} = 2.5$$

(v) **Conclusion:** The calculated value of Z is $|Z| = 2.5$ and the tabulated value of Z is $Z_\alpha = 1.96$ since the calculated value of Z is greater than tabulated value of Z we reject null hypothesis and hence we accept alternative hypothesis, that is, the sample has not been drawn from a normal population with $\mu = 100$.

EXAMPLE 7.2

The income distribution of the population of a certain village has a mean of ₹6,000 and a variance of ₹32,400. Could a sample of 64 persons with a mean income of ₹5,950 belong to this population? Test at 1% level of significance.

Solution:

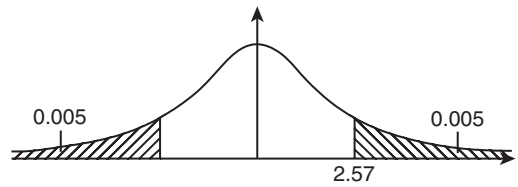
(i) **Null hypothesis:** The sample has been drawn from a normal population with mean ₹5,950

(i.e.,) H_0: $\mu = 6000$

(ii) **Alternative hypothesis:** H_1: $\mu \neq 6000$ (two-tailed test)

(iii) **Level of significance:** $\alpha = 1\%$ level

The critical value of Z at 1% level of significance is $Z_{\frac{\alpha}{2}} = 2.57$.



(iv) **Test statistic:** Under H_0,

$$Z = \frac{\bar{x} - \mu}{\dfrac{\sigma}{\sqrt{n}}}$$

From the given data, $\mu = 6000$, $\sigma = 178.885$ and $\bar{x} = 5950$, $n = 64$

$$\therefore Z = \frac{5950 - 6000}{\dfrac{\sqrt{32000}}{64}} = \frac{-50}{\sqrt{500}} = \frac{-50}{22.3606}$$

$$= -2.23606$$

(v) **Conclusion:** Since the calculated value of Z, $|Z| = 2.23$ is lesser than the critical value of Z, that is, $Z_\alpha = 2.57$, we accept null hypothesis, that is, the sample has been drawn from a population with $\mu = 6000$.

EXAMPLE 7.3

An insurance agent has claimed that the average age of policy holders who insure through him is less than the average for all agents which is 30.5 years. A random sample of 100 policy holders who had insured through him gave the following age distribution:

Age last birthday (years)	16–20	21–25	26–30	31–35	36–40	Total
Number of persons insured	12	22	20	30	16	100

Calculate the arithmetic mean and standard deviation of this distribution and use these values to test his claim at 5% level of significance.

Solution:

C. I.	Mid values x_i	Number of persons insured f_i	$f_i x_i$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})^2 f_i$
16–20	18	12	216	116.64	1399.68
21–25	23	22	506	33.64	740.08
26–30	28	20	560	0.64	12.8
31–35	33	30	990	17.64	529.2
36–0	38	16	608	84.64	1354.24
Total		100	2880		4036

$$\bar{x} = \frac{1}{N}\sum f_i x_i = \frac{2880}{100} = 28.8$$

$$S^2 = \frac{1}{N}\sum f_i (x_i - \bar{x})^2 = \frac{4036}{100} = 40.36,\ S = 6.353$$

(i) **Null hypothesis:** H_0: $\mu = 30.5$ years. The average age of policy holders is 30.5 years.

(ii) **Alternative hypothesis:** H_1: $\mu < 30.5$ years (left-tailed test).

(iii) **Level of significance:** $\alpha = 5\%$ the critical value of Z for 5% level of significance is $Z_\alpha = -1.645$.

(iv) **Test statistic:** Under H_0,

$$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

$$= \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} \qquad \text{Given } \mu = 30.5 \text{ years}$$

$$= \frac{28.8 - 30.5}{\frac{6.353}{\sqrt{100}}}$$

$$Z = \frac{-1.7}{0.6353}$$
$$= -2.676$$

(v) **Conclusion:** Since the calculated value of Z, $|Z| = 2.676$ is greater than the critical value of $Z_\alpha = 1.645$ at 5% level of significance, we reject null hypothesis and accept alternative hypothesis. Hence, the random sample of 100 policy holders who had insured through him has an average of policy holders less than 30.5 years.

EXAMPLE 7.4

The mean life time of a sample of 100 fluorescent light bulbs produced by a company is computed to be 1570 hours with a standard deviation of 120 hours. If μ is the mean life time of all the bulbs produced by the company, test the hypothesis $\mu = 1600$ hours against the alternative hypothesis $\mu \neq 1600$ hours using $\alpha = 5\%$ and 1% levels of significance.

Solution:

(i) **Null hypothesis:** The sample has been drawn from a normal population with mean 1600 hours.

$$H_0: \mu = 1600 \text{ hours.}$$

(ii) **Alternative hypothesis:** $H_1: \mu \neq 1600$ (two-tailed test).

(iii) **Level of significance:** $\alpha = 5\%$ the critical value of Z when $\alpha = 5\%$ is $Z_\alpha = 1.96$.

The critical value of Z when $\alpha = 1\%$ is $Z_\alpha = 2.58$.

(iv) **Test statistic:** Under H_0,

$$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Given that $\mu = 1600$, $\bar{x} = 1570$ hours

$\sigma = 120$ hrs, $n = 100$

$$Z = \frac{1570 - 1600}{\frac{120}{\sqrt{100}}} = \frac{-30}{12} = -2.5$$

(v) **Conclusion:** Since the calculated value $|Z| = 2.5$ is greater than $Z_\alpha = 1.96$, we reject null hypothesis and hence accept alternative hypothesis at 5% level or significance. Hence, the sample has been drawn from a population with $\mu \neq 1600$, at 5% level of significance. Since the calculated value $|Z| = 2.5$ is greater than the critical value $Z_\alpha = 2.58$ we accept null hypothesis. Hence, the sample has been drawn from a normal population.

EXAMPLE 7.5

The mean IQ of a sample of 1600 children was 99. Is it likely that this was a random sample from a population with mean IQ 100 and standard deviation 15?

Solution:

(i) **Null hypothesis:** The mean IQ of the population is 100. $H_0: \mu = 100$

(ii) **Alternative hypothesis:** $\mu \neq 100$ (two-tailed test)

(iii) **Level of significance:** α = 5% level. The critical value of Z, at 5% level of significance $Z_\alpha = 1.96$

(iv) **Test statistic:** Under H_0,

$$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Given that $\bar{x} = 99$, $\mu = 100$, $\alpha = 15$, $n = 1600$

$$Z = \frac{99 - 100}{\frac{15}{\sqrt{1600}}} = \frac{-1}{0.375}$$

$$= -2.666$$

(v) **Conclusion:** Since the calculated value of Z, $|Z| = 2.666$ is greater than the critical value of Z which is $Z_\alpha = 1.96$, we reject null hypothesis at 5% level of significance (or) we accept alternative hypothesis.

Hence the sample is drawn from normal population with $\mu = 100$.

Type II: Testing of Significance for Difference of Means (Large Samples)

Let $\bar{x}_1$ be the mean of a sample of size n_1 from a population with mean μ_1 and variance σ_1^2 and $\bar{x}_2$ be the mean of an independent random sample of size n_2 from a population with mean μ_2 and variance σ_2^2.

Since $\bar{x}_1 \sim N\left(\mu_1, \frac{\sigma_1^2}{n_1}\right)$ and $\bar{x}_2 \sim N\left(\mu_2, \frac{\sigma_2^2}{n_2}\right)$, their difference $\bar{x}_1 - \bar{x}_2$ which is the difference between two normal variates is also a normal variate.

Hence $z = \frac{(\bar{x}_1 - \bar{x}_2) - E(\bar{x}_1 - \bar{x}_2)}{S.E(\bar{x}_1 - \bar{x}_2)} \sim N(0,1)$ where S.E $(\bar{x}_1 - \bar{x}_2)$ is the standard error of the statistic $(\bar{x}_1 - \bar{x}_2)$ (Given in Section 6.5).

Under the null hypothesis, H_0: $\mu_1 = \mu_2$, that is, there is no significant difference between the sample means,

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

Caution:

- If the two samples have been drawn from populations with common standard deviation σ, then $\sigma_1^2 = \sigma_2^2 = \sigma^2$

$$\therefore Z = \frac{\bar{x}_1 - \bar{x}_2}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim N(0,1)$$

- If the common variance, σ^2 is not known, then its estimation based on the sample variances is calculated using,

$$\sigma^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

This formula is used when the sample sizes are not sufficiently large. However, if the sample sizes are sufficiently large, then the common variance can be obtained using,

$$\sigma^2 = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}$$

Worked Out Examples

EXAMPLE 7.6

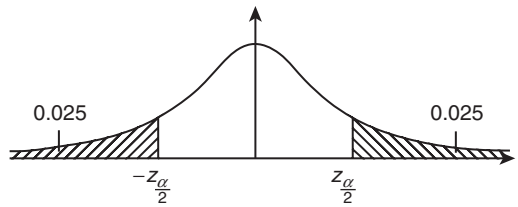
The mean yield of wheat from a district *A* was 210 kg with a standard deviation 10 kg per acre from a sample of 100 plots. In another district *B*, the mean yield was 220 kg with a standard deviation of 12 kg from a sample of 150 plots. Assuming that the standard deviation of the yield in the entire state was 11 kg, test whether there is any significant difference between the mean yields of crops in the two districts.

(i) **Null hypothesis:** There is no significant difference between the mean yield of crops in the two districts, that is, the population means in the two districts are equal. H_0: $\mu_1 = \mu_2$.

(ii) **Alternative hypothesis:** H_1: $\mu_1 \neq \mu_2$ (two-tailed test)

(iii) **Level of significance:** $\alpha = 5\%$

The critical value of Z at 5% level of significance is $Z_{\frac{\alpha}{2}} = 1.96$



(iv) **Test statistic:** The given data is

District A	District B
$x_1 = 100$	$x_2 = 150$
$\bar{x}_1 = 210$	$\bar{x}_2 = 220$
$S_1 = 10$	$S_2 = 12$
$\sigma_1 = 11$	$\sigma_2 = 11$

Under H_0, $Z = \dfrac{\bar{x}_1 - \bar{x}_2}{\sigma\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} \sim N(0,1)$

$$\therefore Z = \frac{210 - 220}{11\sqrt{\dfrac{1}{100} + \dfrac{1}{150}}}$$

$$= \frac{-10}{11\sqrt{\dfrac{5}{300}}} = -7.05$$

Hence, the calculated value of Z, $|Z| = 7.05$.

(v) **Conclusion:** Since the calculated value of $|Z| = 7.05$ is greater than the tabulated value (critical value) of Z that is, $|Z| > Z_\alpha$ we reject null hypothesis at 5% level, that is, we accept alternative

hypothesis at 5% uses of significance. Hence, there is significant difference in the mean yield of crops in the two districts.

EXAMPLE 7.7

The following information is obtained when an intelligence test was given to two groups of boys and girls:

	Mean score	SD	Number
Girls	75	10	50
Boys	70	12	100

Is the difference in the mean score of boys and girls statistically significant?

Solution:

(i) **Null hypothesis:** There is no significant difference between mean score of boys and girls.

$$H_0: \mu_B = \mu_G$$

(ii) **Alternative hypothesis:** $H_1: \mu_B \neq \mu_G$ (two-tailed test)

(iii) **Level of significance:** $\alpha = 5\%$

The critical value of Z for 5% level of significance is $Z_\alpha = 1.96$

(iv) **Test statistic:** Under H_0,

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0.1)$$

Given the data, $\bar{x}_1 = 75$, $\bar{x}_2 = 70$,

$\sigma_1 = 10$, $\sigma_2 = 12$, $n_1 = 50$, $n_2 = 100$

$$Z = \frac{75 - 70}{\sqrt{\frac{10^2}{50} + \frac{12^2}{100}}} = \frac{5}{\sqrt{2 + 1.44}} = \frac{5}{1.854} = 2.696$$

(v) **Conclusion:** Since $|Z| = 2.696$ is very much greater than $Z_\alpha = 1.96$, that is, since the calculated value is greater than tabulated value at 5% level, we reject null hypothesis. Hence, there is a significant difference between the mean score of boys and girls.

EXAMPLE 7.8

A buyer wants to decide which of the two brands of electric bulbs he should buy as he has to buy them in bulk. As a specimen, he buys 100 bulbs of each of the two brands—*A* and *B*. On using these bulbs he finds that brand *A* has a mean life of 1200 hours with a standard deviation of 50 hours and brand *B* has a mean life of 1150 hours with a standard deviation of 40 hours. Do the two brands differ significantly in quality? Test at 1% and 5% levels of significance.

Solution:

(i) **Null hypothesis:** There is no significant difference between the two brands of bulbs in quality.

$$H_0: \mu_A = \mu_B$$

(ii) **Alternative hypothesis:** $H_1: \mu_A \neq \mu_B$

(iii) **Level of significance:** $\alpha = 5\%$ and $\alpha = 1\%$. The critical value of Z at 5% and 1% levels of significance are $Z_{\frac{\alpha}{2}} = 1.96$ and $Z_{\frac{\alpha}{2}} = 2.58$ respectively.

(iv) **Test statistic:** Under H_0,

Given that $\bar{x}_1 = 1200$ hours

$\bar{x}_2 = 1150$ hours

$S_1 = 50$ hours, $S_2 = 40$ hours

$n_1 = 100, n_2 = 100$

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

$$= \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

$$= \frac{1200 - 1150}{\sqrt{\frac{50^2}{100} + \frac{40^2}{100}}} = \frac{50}{\sqrt{25+16}} = \frac{50}{6.403}$$

$$\therefore Z = 7.808$$

(v) **Conclusion:** Since the calculated value of Z is very much greater than the critical values of Z at both 5% and 1%, we reject null hypothesis at both 5% and 1%, that is, the two brands of bulbs differ significantly at 5% and 1% in quality.

EXAMPLE 7.9

A random sample of 100 mill workers at Kanpur showed their mean wage to be ₹3,500 with a standard deviation of ₹280. Another random sample of 150 mill workers in Mumbai showed the mean wage to be ₹3,900 with a standard deviation of ₹400. Do the mean wages of workers in Mumbai and Kanpur differ significantly, at 5% level of significance?

Solution:

(i) **Null hypothesis:** There is no significant difference between the mean wages of workers in Mumbai and Kanpur

$$H_0: \mu_1 = \mu_2$$

(ii) **Alternative hypothesis:** $H_1: \mu_1 \neq \mu_2$ (two-tailed test)

(iii) **Level of significance:** $\alpha = 5\%$. The critical value of Z at 5% level of significance is $Z_\alpha = 1.96$

(iv) **Test statistic:** Under H_0,

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \sim N(0,1)$$

Given that $\bar{x}_1 = 3500 \quad \bar{x}_2 = 3900$

$S_1 = 280 \quad S_2 = 400$

$n_1 = 100 \quad n_2 = 150$

$$Z = \frac{3500 - 3900}{\sqrt{\frac{280^2}{100} + \frac{400^2}{150}}} = \frac{-400}{\sqrt{784 + 1066.6}} = \frac{-400}{43.016}$$

$$Z = -9.298$$

(v) **Conclusion:** Since the calculated value $|Z| = 9.298$ is very much greater than critical value of Z, $Z_\alpha = 1.96$ at 5% level of significance, we reject null hypothesis and hence accept alternative hypothesis. The mean wages of workers in Mumbai and Kanpur are significantly different at 5% level of significance.

EXAMPLE 7.10

On an elementary school examination in spellings, the mean grade of 32 boys was 72 with a standard deviation of 8, while the mean grade of 36 girls was 75 with a standard deviation of 6. Test the hypothesis at 0.01 level of significance that the girls are better in spelling than the boys.

Solution:

(i) **Null hypothesis:** There is no significant difference between the mean grades of boys and girls at a school examination.

$$H_0: \mu_1 = \mu_2$$

(ii) **Alternative hypothesis:** $H_1: \mu_1 \neq \mu_2$ (two-tailed test)

(iii) **Level of significance:** $\alpha = 0.01$. The critical value of z for 1% level of significance is given by $Z_{\frac{\alpha}{2}} = 2.58$.

(iv) **Test statistic:** Under H_0,

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \sim N(0,1)$$

Given that, $\bar{x}_1 = 72, n_1 = 32, S_1 = 8$

$\bar{x}_2 = 75, n_2 = 36, S_2 = 6$

$$\therefore Z = \frac{72 - 75}{\sqrt{\frac{8^2}{32} + \frac{6^2}{36}}} = \frac{-3}{\sqrt{2+1}} = -1.732$$

(v) **Conclusion:** Since the calculated value of Z, $|Z| = 1.732$ is less than the critical value of Z, $Z_{\frac{\alpha}{2}} = 2.58$, we accept null hypothesis. Hence, there is no significant difference between the mean grades of boys and girls at the school examination.

EXAMPLE 7.11

A sample of 100 electric light bulbs produced by manufacturer *A* showed a mean life time of 1190 hours and a standard deviation of 90 hours. A sample of 75 bulbs produced by manufacturer *B* showed a mean life time of 1230 hours with a standard deviation of 120 hours. Test the hypothesis that the bulbs of manufacturer *B* are superior to those of manufacturer *A* using a significant level of (a) 0.05 (b) 0.01.

Solution: Given $\bar{x}_A = 1190$ hours, $n_1 = 100$,

$\bar{S}_A = 90$ hours

$n_2 = 75$, $\bar{x}_B = 1230$ hours, $\bar{S}_B = 120$ hours.

(i) **Null hypothesis:** H_0: $\mu_A = \mu_B$, that is, there is no significant difference between bulbs manufactured by the two manufactures *A* and *B*.

(ii) **Alternative hypothesis:** H_1: $\mu_A < \mu_B$ (left-tailed test)

(iii) **Level of significance:** $\alpha = 5\%$ and $\alpha = 1\%$. The critical value of *Z* for 5% level of significance is $Z_\alpha = 1.645$ and critical value of Z_α for 1% level of significance is $Z_\alpha = 2.33$.

(iv) **Test statistic:** Under H_0, the test statistic is

$$Z = \frac{\bar{x}_A - \bar{x}_B}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{\bar{x}_A - \bar{x}_A}{\sqrt{\frac{S_A^2}{n_1} + \frac{S_B^2}{n_2}}}$$

$$Z = \frac{1190 - 1230}{\sqrt{\frac{90^2}{100} + \frac{120^2}{75}}} = \frac{-40}{\sqrt{81 + 192}} = \frac{-40}{\sqrt{273}} = \frac{-40}{16.5227}$$

$$\therefore Z = -2.421$$

(v) **Conclusion:** Since the calculated value of *Z*, $|Z| = 2.421$ is greater than the critical values of *Z*, $Z_\alpha = 1.645$ and it is also greater than the critical value of *Z*, $Z_\alpha = 2.33$, we reject null hypothesis, and accept alternative hypothesis. Hence, the bulbs of manufacturer *B* are superior to those of manufacturer *A* at 5% and 1% level of significance.

Type III: Test of Significance for Single Proportion

Let us now look at instances where a politician will be interested in knowing what fraction of the voters will favour him in the coming election (or) all manufacturing firms which look at the proportion of defective items when a shipment is made.

Let *X* denotes the number of success in *n* independent trials with constant probability *P* of success for each trial. Then the sample of *n* observations can be treated as a Binomial distribution with $E(X) = nP$ and $V(X) = nPQ$ where $Q = 1 - P$, is the probability of failure. Hence for large *n*, the binomial distribution tends to normal distribution,

$$Z = \frac{x - E(X)}{\sqrt{V(X)}} \sim N(0,1)$$

$$= \frac{X - nP}{\sqrt{nPQ}} \sim N(0,1)$$

If X represents the number of persons who possess a particular attribute (characteristic) then observed proportion of success $P = \frac{X}{n}$

$$\therefore Z = \frac{P - E(p)}{\sqrt{V(p)}} = \frac{p - P}{\sqrt{\frac{PQ}{n}}}$$

where P is the population proportion.

Caution:

If X is the number of persons possessing an attribute then

$$E(p) = E\left(\frac{x}{n}\right) = \frac{1}{n}E(x)$$

$$= \frac{1}{n}(nP) = P$$

and

$$\text{var}\,(p) = V\left(\frac{x}{n}\right) = \frac{1}{n^2}V(x)$$

$$= \frac{1}{n^2}(nPQ)$$

$$= P\frac{Q}{n}.$$

Worked Out Examples

EXAMPLE 7.12

The manufacturer of a patent medicine claimed that it was 90% effective in relieving an allergy for a period of 8 hours. In a sample of 200 people who had the allergy, the medicine provided relief for 160 people. Determine whether the manufacturer's claim is legitimate at 0.01 level of significance.

Solution:

(i) **Null hypothesis:** H_0: $P = 90\% = 0.9$.
It was 90% effective in relieving an allergy for a period of 8 hours.

(ii) **Alternative hypothesis:** H_0: $P \neq 0.9$ (two-tailed test)

(iii) **Level of significance:** $\alpha = 5\%$. The critical value of Z for 5% level of significance is

$$Z_\alpha = 2.50.$$

(iv) **Test statistic:** Under H_0, the test statistic is

$$Z = \frac{p - P}{\frac{PQ}{n}}, \qquad p = \frac{X}{n}$$

where, X = number of persons who got relief from medicine = 160

n = number of people in the sample = 200

$\therefore p = 0.8$, $P = 0.9$

$$Q = 1 - P = 0.1$$

$$\therefore\ Z = \frac{0.8 - 0.9}{\sqrt{\frac{(0.9)(0.1)}{200}}} = \frac{-0.1}{\sqrt{0.00045}}$$

$$= -4.714$$

(v) **Conclusion:** Since calculated value of Z, $|Z| = 4.714$ is greater than critical value, $Z_\alpha = 2.58$, we reject null hypothesis and accept alternative hypothesis. Hence the manufacturer's claim is not true.

EXAMPLE 7.13

It is observed in a survey that 15% of households in a certain city indicated that they owned a washing machine. The survey based on a random sample of 900 households was taken and it was found that 189 households had a washing machine. Can we conclude that there has been a significant increase in the sale of washing machines at 5% level of significance?

Solution:

(i) **Null hypothesis:** There has been no significant increase in the sale of washing machines at 5% level of significance. $P = 15\%$.

(ii) **Alternative hypothesis:** $P > 15\%$ (right-tailed test)

(iii) **Level of significance:** $\alpha = 5\%$. The critical value of Z at 5% level of significance is

$$Z_\alpha = 1.645.$$

(iv) **Test statistic:** Under H_0, the test statistic is

$$Z = \frac{p - P}{\sqrt{\frac{PQ}{n}}}, \qquad p = \frac{X}{n}$$

where, X = households with a washing machine = 189

n = sample of households = 900

$$p = \frac{189}{900} = 0.21,\ P = 0.15,\ Q = 0.85$$

$$Z = \frac{0.21 - 0.15}{\sqrt{0.85 \times \frac{0.15}{900}}} = \frac{0.06}{0.0119} = 5.042$$

(v) **Conclusion:** Since the calculated value of Z, $|Z| = 5.041$ is greater than the critical value of Z, $Z_\alpha = 1.645$, we reject null hypothesis and accept alternative hypothesis.

Hence, there has been a significant increase in the sale of washing machines at 5% level of significance.

EXAMPLE 7.14

An advertisement company feels that 20% of the population in the age group of 18 to 25 years in a town watches a particular serial. To test this assumption, a random sample of 2,000 individuals in the same age group was taken of which 440 watched the serial. At 5% level of significance, can we accept the assumption laid down by the company?

Solution:

(i) **Null hypothesis:** $P = 20\%$. Advertisement company feels that 20% of the population in the age group of 18 to 25 years in a town watch a particular serial.

(ii) **Alternative hypothesis:** $P \neq 20\%$ (two-tailed test)

(iii) **Level of significance:** $\alpha = 5\%$. The critical value of Z at 5% level of significance is

$$Z_{\alpha} = 1.96.$$

(iv) **Test statistic:** Under H_0, the test statistic is

$$Z = \frac{p - P}{\sqrt{\frac{PQ}{n}}}, \qquad p = \frac{X}{n}$$

$X = 440$ individual who watched the serial

$n = 2000$, sample of individuals in the age group of 18 to 25 years.

$$p = \frac{440}{2000} = 0.22$$

$P = 0.2$, $Q = 0.8$

$$Z = \frac{0.22 - 0.2}{\sqrt{\frac{0.2 \times 0.8}{2000}}} = \frac{0.02}{0.0089} = 2.236$$

(v) **Conclusion:** Since the calculated value of Z, $|Z| = 2.236$ is greater than the critical value of Z, $Z_{\alpha} = 1.96$, we reject null hypothesis and accept alternative hypothesis. Hence, the assumption of advertisement company that 20% of population watch the serial cannot be accepted.

EXAMPLE 7.15

A dice is thrown 49,152 times and of these 25,145 yielded either 4 or 5 or 6. Is this consistent with the hypothesis that the dice must be unbiased?

Solution:

(i) **Null hypothesis:** Dice is unbiased.

$$H_0: P = \frac{1}{2} = 0.5$$

(ii) **Alternative hypothesis:** $H_1: P \neq \frac{1}{2}$ (two-tailed test)

(iii) **Level of significance:** $\alpha = 5\% = 0.05$. The critical value of Z at 5% level of significance is $Z_{\frac{\alpha}{2}} = 1.96$.

(iv) **Test statistic:** Under H_0,

$$Z = \frac{p - P}{\sqrt{\frac{PQ}{n}}}$$

If occurrence of 4 or 5 or 6 is termed as success then we are given that

p = proportion of success

$$p = \frac{X}{n}$$

$$= \frac{25145}{49152} = 0.512$$

$P = 0.5, Q = 0.5$

$$\therefore Z = \frac{0.512 - 0.5}{\sqrt{\frac{(0.5)(0.5)}{49152}}}$$

$$= \frac{0.012 \times 221.70}{0.5}$$

$$|z| = 5.32$$

(v) **Conclusion:** The calculated value of Z is $|Z| = 5.32$ and the tabulated value of Z at 5% level of significance is $Z_\alpha = 1.96$, since the calculated value of Z is greater than the tabulated value of Z at 5% level, we reject the null hypothesis and accept the alternative hypothesis. Hence, the dice is unbiased.

EXAMPLE 7.16

An automobile manufacturer asserts that the seat belts of his seats are 90% effective. A consumer group tests the seat belts on 50 cars and finds it effective on 37 of them. Test the collection of manufacturer assertion at 5% level based on the observed data.

Solution:

(i) **Null hypothesis:** $P = 90\%$ the seat belts of his seats of an automobile manufacturing are 90% effective.

(ii) **Alternative hypothesis:** P ≠ 90% (two-tailed test)

(iii) **Level of significance:** $\alpha = 5\%$. The critical value of Z at 5% level of significance is $Z_\alpha = 1.96$.

(iv) **Test statistic:** Under H_0, the test statistic is

$$Z = \frac{p - P}{\sqrt{\frac{PQ}{n}}}, \; p = \frac{X}{n}$$

where $X = 37$, number of cars with effective seat belts

$n = 50$, consumer group on which the test is made

$$p = \frac{37}{50} = 0.74$$

$P = 0.9, Q = 0.1$

$$Z = \frac{0.74 - 0.9}{\sqrt{\frac{0.9 \times 0.1}{50}}} = \frac{-0.16}{0.042} = -3.77$$

(v) **Conclusion:** Since the calculated value of Z, $|Z| = 3.77$ is greater than critical value of Z, $Z_\alpha = 1.96$, we reject null hypothesis and accept alternative hypothesis. Hence, the automobile manufacturer's assertion that the seat belts of his seats are not 90% effective.

Type IV: Test of Significance for Difference of Proportions

Extending the theory that was studied in Type III, let X_1 and X_2 denote the number of persons possessing a particular attribute (characteristic) in random samples of sizes n_1 and n_2 from two populations, respectively. Then the sample proportions are given by $p_1 = \frac{x_1}{n_1}$ and $p_2 = \frac{x_2}{n_2}$ then $E(p_1) = P_1$, $E(p_2) = P_2$.

Where p_1 and p_2 are population proportions and $v(p_1) = \frac{P_1Q_1}{n_1}$ $v(p_2) = \frac{P_2Q_2}{n_2}$ where $Q_1 = 1 - P_1$, $Q_2 = 1 - P_2$.

Since p_1 and p_2 are independent and normally distributed, their difference $(p_1 - p_2)$ is also normally distributed.

$$\therefore Z = \frac{(p_1 - p_2) - E(p_1 - p_2)}{\sqrt{V(p_1 - p_2)}} \sim N(0,1)$$

Under the null hypothesis that there is no significant difference between the sample proportion, i.e., H_0: $P_1 = P_2$ we have $E(p_1 - p_2) = 0$,

$$V(p_1 - p_2) = PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$$

Since $P_1 = P_2 = P$, $\quad Q_1 = Q_2 = Q$

$$\therefore Z = \frac{p_1 - p_2}{\sqrt{PQ\left(\frac{1}{n_2} + \frac{1}{n_2}\right)}} \sim N(0,1)$$

Caution:

- Suppose the population proportions P_1 and P_2 are given to be different, that is, $P_1 \neq P_2$ and if we want to test if the difference $(P_1 - P_2)$ in population proportions is likely to be hidden in samples of sizes of n_1 and n_2 from two populations, respectively.

  Then $$Z = \frac{(p_1 - p_2) - E(P_1 - P_2)}{\sqrt{\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}}} \sim N(0,1)$$

- In case the sample proportions are not given and if we want to test if the difference in population proportions is likely to be hidden in sampling, the test statistic is

$$Z = \frac{|p_1 - p_2|}{\sqrt{\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}}} \sim N(0,1).$$

- Suppose we do not have any information of the population proportions then, under H_0: $P_1 = P_2 = P$, an estimate of P based on the two random samples is obtained as

$$\hat{P} = \frac{n_1p_1 + n_2p_2}{n_1 + n_2} = \frac{X_1 + X_2}{n_1 + n_2}$$

- Suppose we want to test the significance of the difference between p_1 and $p = \dfrac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$

  then $$Z = \frac{p_1 - p}{\sqrt{\dfrac{n_2}{n_1 + n_2} \cdot \dfrac{pq}{n_1}}} \sim N(0,1)$$

Worked Out Examples

EXAMPLE 7.17

In a sample of 600 men from a certain city, 450 men are found to be smokers. In a sample of 900 men from another city, 450 are found to be smokers. Does the date indicate that the two cities are significantly different with respect to prevalence of smoking habit among men?

Solution:

(i) **Null hypothesis:** H_0: $P_1 = P_2$. There is no significant difference in the two cities with respect to prevalence of smoking habit among men.

(ii) **Alternative hypothesis:** H_1: $P_1 \neq P_2$ (two-tailed test)

(iii) **Level of significance:** $\alpha = 5\%$ the critical value of Z at 5% level of significance is

$$Z_\alpha = 1.96.$$

(iv) **Test statistic:** Under H_0, the test statistic is

$$\therefore Z = \frac{p_1 - p_2}{\sqrt{PQ\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$$

$$p_1 = \frac{X_1}{n_1},$$

X_1 = number of smokers in city $A = 450$

n_1 = number of men considered in the sample = 600

$$p_1 = \frac{450}{600} = 0.75$$

$$p_2 = \frac{X_2}{n_2},$$

X_2 = number of smokers in city $B = 450$

n_2 = number of men considered in the sample = 900

$$p_2 = \frac{450}{900} = 0.5$$

P = % smoking habit in city A or $B = 0.5$, $Q = 0.5$

$$Z = \frac{0.75 - 0.5}{\sqrt{0.5 \times 0.5\left(\dfrac{1}{600} + \dfrac{1}{900}\right)}} = \frac{0.25}{0.026} = 9.486$$

(v) **Conclusion:** Since $|Z| > Z_\alpha$, we reject H_0 and accept H_1. Hence the two cities are significantly different with respect to prevalence of smoking habit, among men.

EXAMPLE 7.18

A sample survey results show that out of 800 literate people, 480 are employed whereas out of 700 illiterate people only 350 are employed. Can the difference between two proportions of employed persons be described due to sampling fluctuations?

Solution:

(i) **Null hypothesis:** H_0: $P_1 = P_2$ The difference between two proportions of employed persons be attributed to sampling fluctuations.

(ii) **Alternative hypothesis:** H_1: $P_1 \neq P_2$ (Two-tailed test)

(iii) **Level of significance:** $\alpha = 5\%$. The critical value of Z at 5% level of significance, $Z_\alpha = 1.96$.

(iv) **Test statistic:** Under H_0, the test statistic is

$$\therefore Z = \frac{p_1 - p_2}{\sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$p_1 = \frac{X_1}{n_1},$$

$X_1 = 480$, number of employed people out of literates

n_1 = number of literate people in the survey = 800

$$p_1 = \frac{480}{800} = 0.6$$

$$p_2 = \frac{X_2}{n_2},$$

X_2 = number of employed people out of illiterate people = 350

n_2 = number of illiterate people in the sample = 700

$$p_2 = \frac{350}{700} = 0.5$$

P = % proportion of employed people = 0.5, $Q = 0.5$

$$Z = \frac{0.6 - 0.5}{\sqrt{0.5 \times 0.5\left(\frac{1}{800} + \frac{1}{700}\right)}} = \frac{0.1}{0.025} = 3.864$$

(v) **Conclusion:** Since $|Z| = Z_\alpha$ at 5% level, we reject H_0 and accept H_1. Hence the two proportions of employed persons are significantly different.

EXAMPLE 7.19

In a large city *A,* 25% of a random sample of 900 school boys had defective eye sight. In another large city *B*, 15.5% of a random sample of 1,600 school boys had the same defect. Is the difference between the two proportions significant?

Solution:

(i) **Null hypothesis:** H_0: $P_1 = P_2$. The two proportions of school boys in the two cites A and B having defective eye sight are not significantly different.

(ii) **Alternative hypothesis:** H_1: $P_1 \neq P_2$ (Two-tailed test)

(iii) **Level of significance:** $\alpha = 5\%$ the critical value of z at 5% level of significance, $Z_\alpha = 1.96$

(iv) **Test statistic:** Under H_0, the test statistic is

$$\therefore Z = \frac{p_1 - p_2}{\sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$p_1 = \frac{X_1}{n_1},$$

X_1 = number of students with defective eye sight from city $A = 25\%$

$n_1 = 900$, schools boys of city A

$p_1 = 0.25$

$$p_2 = \frac{X_2}{n_2},$$

X_2 = number of school boys with defective eye sight from city $B = 15.5\%$

$n_2 = 1600$, schools boys of city B

$p_2 = 0.155$

P = Proportion of boys of the two cities with defective eye sight = 0.5

$$Z = \frac{0.25 - 0.155}{\sqrt{0.5 \times 0.5\left(\frac{1}{900} + \frac{1}{1600}\right)}} = \frac{0.095}{0.0208} = 4.56$$

(v) **Conclusion:** Since calculated $|Z|$ > critical value Z, we reject null hypothesis, and accept alternative hypothesis. Hence, the proportions of school boys of the two cities with defective eye sight are significantly different.

EXAMPLE 7.20

In a winter of an epidemic flu, 2000 babies were surveyed, by a well-known pharmaceutical company to determine if the company's new medicine was effective after two days. Among 120 babies, who had the flu and were given the medicine 29 were cured within 2 days. Among 280 babies who had the flu but were not given the medicine, 56 were cured within 2 days. Is there any significant indication that supports the company's claim of the effectiveness of the medicine?

Solution:

(i) **Null hypothesis:** $P_1 = P_2$. There is no significant indication that supports the company's claim of the effectiveness of the medicine.

(ii) **Alternative hypothesis:** $P_1 \neq P_2$ (two-tailed test)

(iii) **Level of significance:** $\alpha = 5\%$. The critical value of Z, at 5% level of significance is

$$Z_\alpha = 1.96.$$

(iv) **Test statistic:** Under H_0, the test statistics is

$$Z = \frac{p_1 - p_2}{\sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$p_1 = \frac{X_1}{n_1} = \frac{29}{120} = 0.2416$$

X_1 = Babies who were given medicine and cured from flu

n_1 = number of babies who had flu

$$p_2 = \frac{X_2}{n_2} = \frac{56}{280} = 0.2$$

X_2 = babies who had flu and were cured without medicine

n_2 = number of babies who had flu

P = proportion of persons with medicine = 0.5

Q = proportion of persons without medicine = 0.5

$$Z = \frac{0.2416 - 0.2}{\sqrt{0.5 \times 0.5\left(\frac{1}{120} + \frac{1}{280}\right)}} = \frac{0.0416}{0.0545} = 0.76$$

(v) **Conclusion:** Since $|Z| = 0.76$ is less than $Z_\alpha = 1.96$ at 5%, we accept null hypothesis. Hence there is no significant indication that supports the company's claim of the effectiveness of the medicine.

EXAMPLE 7.21

In a study to estimate the proportion of residents in a certain city and its suburban residents who favour the construction of a nuclear power plant, it is found that 63 out of 100 urban residents favour the construction, while only 59 out of 125 suburban residents favour the construction of plant. Is there a significant difference between the proportion of urban and suburban residents who favour the construction of nuclear power plant?

Solution:

(i) **Null hypothesis:** H_0: $P_1 = P_2$. There is no significant difference between the proportion of urban and suburban residents who favour the construction of nuclear power plant.

(ii) **Alternative hypothesis:** H_1: $P_1 \neq P_2$ (two-tailed test)

(iii) **Level of significant:** $\alpha = 5\%$. The critical value of Z at 5% level of significance is $Z_\alpha = 1.96$.

(iv) **Test statistic:** Under H_0, the test statistic is

$$Z = \frac{p_1 - p_2}{\sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}},$$

$$p_1 = \frac{X_1}{n_1} = \frac{63}{100} = 0.63$$

where, X_1 = number of urban residents favouring the construction of nuclear power plant

n_1 = number of urban residents

$$p_2 = \frac{X_2}{n_2} = \frac{59}{125} = 0.475$$

X_2 = number of suburban residents favouring the construction of nuclear power plant

n_2 = number of suburban residents

P = proportion of urban and suburban residents who favour the construction of nuclear power plant

$$Z = \frac{0.63 - 0.475}{\sqrt{0.5 \times 0.5\left(\frac{1}{100} + \frac{1}{125}\right)}} = \frac{0.155}{0.06708} = 2.31$$

(v) **Conclusion:** Since $|Z| = 2.31$ is greater than, the critical value of Z, $Z_\alpha = 1.96$, we reject null hypothesis and accept alternative hypothesis. Hence, the proportion of urban and suburban residents who favour the construction of nuclear power plant are not equal.

Type V: Test of Significance for Difference of Two Standard Deviations

Let us now consider whether two samples having different standard deviations come from the same population. Since we usually do not know the standard deviations of the population, we have to resist to standard deviations of the samples provided, the samples are very large.

Under null hypothesis that the sample standard deviation do not differ significantly, that is, H_0: $\sigma_1 = \sigma_2$

$$Z = \frac{S_1 - S_2}{\sqrt{\frac{\sigma_1^2}{2n_1} + \frac{\sigma_2^2}{2n_2}}} \sim N(0,1)$$

where S_1 and S_2 are sample standard deviations and $E(S_1 - S_2) = 0$ and $V(S_1 - S_2) = \frac{\sigma_1^2}{2n_1} + \frac{\sigma_2^2}{2n_2}$

Worked Out Examples

EXAMPLE 7.22

A large organization produces electric bulbs in each of its two factories. The efficiency in the two factories is not alike. So a test is carried out by ascertaining the variability of the life of the bulbs produced by each factory. The results are as follows:

	Factory A	Factory B
Number of bulbs in sample	100(n_1)	200(n_1)
Average life	1100($\bar{X}_1$)	900($\bar{X}_2$)
Standard deviation	240(S_1)	220(S_2)

Determine whether the difference between the variability of life of bulbs from each sample is significant at 1% level of significance.

Solution:

(i) **Null hypothesis:** H_0: $\sigma_1 = \sigma_2$. There is no significant difference between the variability of life of bulbs from each of the samples from the factories *A* and *B*.

(ii) **Alternative hypothesis:** H_1: $\sigma_1 \neq \sigma_2$ (two-tailed test)

(iii) **Level of significance:** $\alpha = 5\%$. The critical value of *Z* for 1% level of significance

$$Z_\alpha = 2.58.$$

(iv) **Test statistic:** Under H_0, the test statistic is

$$Z = \frac{S_1 - S_2}{\sqrt{\frac{\sigma_1^2}{2n_1} + \frac{\sigma_2^2}{2n_2}}} = \frac{S_1 - S_2}{\sqrt{\frac{S_1^2}{2n_1} + \frac{S_2^2}{2n_2}}}$$

$$Z = \frac{240 - 220}{\sqrt{\frac{240^2}{2(100)} + \frac{220^2}{2(200)}}} = \frac{20}{20.223} = 0.989$$

(v) **Conclusion:** Since the calculated $Z = 0.989$ is less than the critical value of *Z*, $Z_\alpha = 2.58$, we accept null hypothesis. Hence, the different between variability of life of bulbs from each sample from two factories *A* and *B* is not significant.

EXAMPLE 7.23

Random samples drawn from two countries gave the following data relating to the height of adult males:

	Country A	Country B
Mean height (in inches)	67.42($\bar{X}_1$)	67.25($\bar{X}_2$)
Standard deviation (in inches)	2.58(S_1)	2.50(S_2)
Number in samples	1000(n_1)	1200(n_2)

Is there a significant difference between the standard deviations?

Solution:

(i) **Null hypothesis:** H_0: $\sigma_1 = \sigma_2$. There is no significant difference between the standard deviations of random samples drawn for two countries relating to the height of adult males.

(ii) **Alternative hypothesis:** H_1: $\sigma_1 \neq \sigma_2$ (two-tailed test)

(iii) **Level of significance:** $\alpha = 5\%$. The critical value of *Z* at 5% level of significance, is $Z_\alpha = 1.96$.

(iv) **Test statistic:** Under H_0, the test statistic is

$$Z = \frac{S_1 - S_2}{\sqrt{\frac{\sigma_1^2}{2n_1} + \frac{\sigma_2^2}{2n_2}}} = \frac{2.58 - 2.50}{\sqrt{\frac{2.58^2}{2(1000)} + \frac{2.50^2}{2(1200)}}}$$

$$Z = \frac{0.08}{0.07702} = 1.0386$$

(v) **Conclusion:** Since calculated Z, $|Z| = 1.0386$ is less than critical value of Z, $Z_\alpha = 1.96$, we accept null hypothesis. Hence there is no significant difference between the standard deviations of random samples drawn from two countries relating to the height of adult males.

Work Book Exercises

1. Past experience indicates that the time for senior school students to complete a standardized test is a normal random variable with a mean of 35 minutes. If a random sample of 50 seniors took an average of 33.1 minutes to complete this test with a standard deviation of 4.3 minutes, test the hypothesis at the 0.05 level of significance that $\mu = 35$ minutes against the alternative that $\mu =$ 35 minutes.
2. A company is engaged in the packaging of superior quality tea jars of 500 gm each. The company is of the view that as long as jars contain 500 gm of tea, the process is in control. The standard deviation is 50 gm. A sample of 225 jars is taken at random and the sample average is found to be 510 gm. Has the process gone out of control?
3. A radio shop sells on an average, 200 radios per day with a standard deviation of 50 radios. After an extensive advertising campaign, the management will compute the average sales for the radio shop next 25 days to see whether an improvement has occurred. Assume that the daily sales of radios are normally distributed. Test the hypothesis at 5% level of significance if $\bar{x} = 216$.
4. A specified brand of automobile tyre is known to average 10,000 km with a standard deviation of 500 km. A random sample of 100 tyres of this brand when tested, resulted in the average of 9,900 km. Is there any evidence to conclude that the quality with respect to this characteristic of this brand of tyres has shipped down, at 1% level of significance?

[**Ans.:** $Z = -2$]

5. A random sample of 200 tins of coconut oil gave an average weight of 4.95 kgs with a standard deviation of 0.21. Do we accept the hypothesis of net weight of 5 kgs per tin at 1% level of significance?

[**Ans.:** $Z = -3.3$]

6. The following information is given relating to two places *A* and *B*. Test whether there is a significant difference between their mean wages.

	A	*B*
Mean wages (₹)	47	49
Standard deviation (₹)	28	40
Number of workers	1000	1500

[**Ans.:** $Z = -1.81$]

7. The following table presents data on the values of a harvested crop stored outside and inside a go down?

	Sample size	Mean	$\sum(x_i - \bar{x})^2$
Outside	40	117	8,685
Inside	100	132	27,315

Assuming that the two samples are random and they have been drawn from normal populations with equal variances, examine if the mean value of the harvested crop is affected by weather conditions.

8. An examination was given to two classes consisting of 40 and 50 students, respectively. In the first class, the mean grade was 74 with a standard deviation of 8, while in the second class the mean grade was 78 with a standard deviation of 7. Is there a significant difference between the performances of the two classes at a level of 0.01 significance?

9. To test the effects of a new fertilizer on wheat production, a tract of land was divided into 60 squares of equal areas, all portions having identical qualities as to soil, exposure to sunlight, etc. The new fertilizer was applied to 30 squares, and the old fertilizer was applied to remaining squares. The mean number of bushels of wheat harvested per square of land using the new fertilizer was 18.2 with a standard deviation of 0.63 bushels. The corresponding mean and standard deviation for the squares using the old fertilizer were 17.8 and 0.54 bushels, respectively. Test at 0.05 and 0.01 levels of significance that new fertilizer is better than the old one.

10. A manufacturer claims that the average tensile strength of thread *A* exceeds the average tensile strength of thread *B* by at least 12 kg. To test his claim, 50 pieces of each type of thread are tested under similar conditions. Type *A* thread has an average tensile strength of 86.7 kg with a standard deviation of 6.28 kg, while type *B* thread had an average tensile strength of 77.8 kg with a standard deviation of 5.61 kg. Test the manufacturer's claim using a 0.05 level of significance.

11. A random sample of size $n_1 = 25$ taken from a normal population with a standard deviation $\sigma_1 = 5.2$, has a mean $\bar{x}_1 = 81$. A second random sample of size $n_2 = 36$ taken from a different normal population with a standard deviation $\sigma_2 = 3.4$, has a mean $\bar{x}_2 = 76$. Test the hypothesis that $\mu_1 = \mu_2$ against the alternative $\mu_1 \neq \mu_2$.

12. A fuel oil company claims that one-fifth of the homes in a certain city are heated by oil. Do we have reason to doubt this claim, if in a random sample of 1,000 homes in this city, it is found that 136 are heated by oil?

13. A coin is tossed 20 times, resulting in 5 heads. Is this sufficient evidence to reject the hypothesis that the coin is balanced in favour of the alternative that heads occur less than 50% of the time?

14. A company engaged in the manufacture of superior quality diaries, which are primarily meant for senior executives in the corporate world. It claims that 75% of the executives employed in Delhi use its diaries. A random sample of 800 executives was taken and it was found that 570 executives did use its diary when the survey was undertaken. Verify the company's claim using 5% level of significance.

[**Ans.:** $Z = -2.45$]

15. In a sample of 500 people in Kerala, 280 are tea drinkers and the rest are coffee drinkers. Can we assume that both coffee and tea are equally popular in this state at 1% level of significance?

[**Ans.:** $Z = 2.68$]

16. A manufacturer claimed that at least 95% of the equipment which he supplied to a factory conformed to specifications. An examination of a sample of 200 pieces of equipment revealed that 18 were faulty. Test his claim at 0.05 and 0.01 level of significance.

17. At a certain date in a large city 16 out of a random sample of 500 men were found to be drinkers. After the heavy increase in tax on intoxicants another random sample of 100 men in the same city included 3 drinkers. Was the observed decrease in the proportion of drinkers significant?

[**Ans.:** $Z = 1.04$]

18. A machine produced 20 defective articles in a batch of 400. After overhauling it produced 10 defective articles in a batch of 300. Has the machine improved? Test at 1% level of significance.
19. Two urns *A* and *B* contain equal number of marbles, but the proportion of red and white marbles in each of the urns is unknown. A sample of 50 marbles selected with replacement from each of the urns revealed 32 red marbles from *A* and 23 red marbles from *B*. Test the hypothesis that the urn *A* has a greater proportion of red marbles than *B*.

[**Ans.:** $Z = 1.818$]

20. The cinema-goers were 800 persons out of a sample of 1,000 persons during the period of a fortnight in a town where no TV programme was available. The cinema-goers were 700 people out of a sample of 2,800 persons during a fortnight in another town where TV programme was available. Do you think there has been a significant decrease in proportion of cinema-goers after introducing TV sets? Test the hypothesis that there is no difference in proportion of cinema-goers using a 0.01 level of significance.
21. A candidate for election made a speech in city *A* but not in *B*. A sample of 500 voters from city *B* showed that 59.6% of the voters were in favour of him, whereas a sample of 300 voters from city *B* showed that 50% of the voters favoured him. Use 5% level of significance to test whether his speech could produce any effect on voters in city *A*.
22. Test the hypothesis that the average content of containers of a particular lubricant is 10 litres if the contents of a random sample of 10 containers are 10.2, 9.7, 10.1, 10.3, 10.1, 9.8, 9.9, 10.4, 10.3, and 9.8 litres. Use 1% level of significance and assume that the distribution of contents is normal.
23. The mean yield of two sets of plots and their variability are given. Examine the following:
    (i) Whether the difference in the mean yield of two sets of plots is significant
    (ii) Whether the difference in the variability in yields is significant.

	Set of 40 plots	Set of 60 plots
Mean yield per plot	1,258 kg	1243 kg
SD per plot	34 kg	28 kg

DEFINITIONS AT A GLANCE

Statistical Hypothesis: It is an assertion or conjunction or a statement concerning one or more populations.

Null Hypothesis: A definite statement we wish to test about the population parameter, that is, a hypothesis of no difference (H_0).

Alternative Hypothesis: A statement complementary to null hypothesis (H_1).

Type I Error: Reject null hypothesis when it is true.

Type II Error: Accept null hypothesis when H_1 is true.

Level of Significance: The maximum probability of making Type I error.

Critical Region: The region of the standard normal curve corresponding to predetermined level of significance for knowing the probability of making Type I error.

Critical Value: The value of the test statistic which separates the critical region and the acceptance region.

One-tailed test: A test of statistical hypothesis where the alternative hypothesis is one-tailed (right or left-tailed).

Two-tailed test: A test of statistical hypothesis where the alternative is two-tailed.

Decision rule(conclusion of a testing procedure): If the calculated test statistic falls within the critical region null hypothesis is rejected, otherwise not.

Beta(β): The probability of not rejecting a null hypothesis when it is false. It is probability of committing Type II error.

FORMULAE AT A GLANCE

- Z-test for testing the significance of a single mean

  $Z = \dfrac{\bar{x} - \mu}{\dfrac{\sigma}{\sqrt{n}}}$ or when σ is not

  Known, $Z = \dfrac{\bar{x} - \mu}{\dfrac{S}{\sqrt{n}}}$ where $S = \dfrac{1}{n-1}\sum (x_i - \bar{x})^2$

- Z-test for testing the significance of difference of means

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$$

  If $\sigma_1^2 = \sigma_2^2 = \sigma^2$ then $z = \dfrac{\bar{x}_1 - \bar{x}_2}{\sigma\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$.

- If sample variances are known, then $\sigma^2 = \dfrac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2 - 2}$

- Z-test for single proportion,

  $Z = \dfrac{p - P}{\sqrt{\dfrac{PQ}{n}}} = \dfrac{X - np}{\sqrt{nPQ}}$ where $p = \dfrac{X}{n}$

- Z-test the difference of two proportions, $Z = \dfrac{p_1 - p_2}{\sqrt{PQ\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$

- Suppose X_1, X_2, n_1, and n_2 are known then common variance $\hat{p} = \dfrac{X_1 + X_2}{n_1 + n_2}$ instead of P.
- Suppose we want to test the difference between p_1 and $p = \dfrac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$ then $Z = \dfrac{p_1 - p}{\sqrt{\dfrac{n_2}{n_1 + n_2} \cdot \dfrac{pq}{n_1}}}$
- Z-test for difference of two standard deviations $Z = \dfrac{S_1 - S_2}{\sqrt{\dfrac{\sigma_1^2}{2n_1} + \dfrac{\sigma_2^2}{2n_2}}}$

OBJECTIVE TYPE QUESTIONS

1. P{Reject H_0 when it is true} is given by
   (a) level of significance α (b) β
   (c) Type I error (d) Type II error
2. If $H_0: \mu = \mu_0$, $H_1: \mu > \mu_0$ is called
   (a) two-tailed test (b) left-tailed test
   (c) right-tailed test (d) none
3. Accepting null hypothesis when it is false or when alternative hypothesis is true is called
   (a) Type I error (b) Type II error
   (c) Type III error (d) none
4. The critical value of the test statistic depends upon which one of the following?
   (a) Level of significance
   (b) Alternative hypothesis whether one-tailed & two-tailed
   (c) Both (a) and (b)
   (d) None
5. The test statistic used to test for single mean for large samples, is $Z =$
   (a) $\dfrac{\bar{x} - \mu}{\dfrac{\sigma}{n}}$ (b) $\dfrac{\bar{x} - \mu}{\dfrac{\sigma}{\sqrt{n}}}$
   (c) $\dfrac{\bar{x} - \sigma}{\dfrac{\mu}{\sqrt{n}}}$ (d) none
6. The common variance used in testing for difference of two means is $\sigma^2 =$
   (a) $\dfrac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2}$ (b) $\dfrac{n_1 s_1 + n_2 s_2}{n_1 + n_2 - 2}$
   (c) $\dfrac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}$ (d) none

7. The test statistic used for testing difference of means for large samples is $Z =$

(a) $\dfrac{\bar{x}_1 - \bar{x}_2}{\dfrac{\sigma_1}{n_1} + \dfrac{\sigma_2}{n_2}}$ (b) $\dfrac{\bar{x}_1 - \bar{x}_2}{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}$

(c) $\dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$ (d) none

8. If X represents the number of persons who possess a particular attribute or characteristic from among a sample of n persons then observed proportion of success is $P =$ _____.

(a) $\dfrac{X}{n}$ (b) $X \cdot n$

(c) $n - X$ (d) none

9. The expectation and variance of observed proportion of success are given by

(a) $E(p) = P$, $V(p) = PQ$ (b) $E(p) = P$, $V(p) = \dfrac{PQ}{n}$

(c) $E(p) = P$, $V(p) = \dfrac{PQ}{\sqrt{n}}$ (d) none

10. The test statistic used to test for single proportion of success is $Z =$ _____.

(a) $\dfrac{p - P}{\dfrac{PQ}{n}}$ (b) $\dfrac{p - P}{\sqrt{\dfrac{PQ}{n}}}$

(c) $\dfrac{p - P}{\dfrac{PQ}{\sqrt{n}}}$ (d) none

11. Under H_0: $P_1 = P_2 = P$, an estimate for population proportion P is given by $\hat{p} =$_____.

(a) $n_1 p_1 + n_2 p_2$ (b) $\dfrac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$

(c) $\dfrac{p_1 + p_2}{n_1 + n_2}$ (d) none

12. The test statistic used for testing the difference of two proportions is $Z =$ _____.

(a) $\dfrac{p_1 - p_2}{\sqrt{\dfrac{PQ}{n_1 + n_2}}}$ (b) $\dfrac{p_1 - p_2}{\sqrt{PQ(n_1 + n_2)}}$

(c) $\dfrac{p_1 - p_2}{\sqrt{PQ\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$ (d) none

13. The test statistic used to test H_0: $\sigma_1 = \sigma_2$ against H_1: $\sigma_1 \neq \sigma_2$ is $Z =$ _____.

(a) $\dfrac{S_1 - S_2}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$

(b) $\dfrac{S_1 - S_2}{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}$

(c) $\dfrac{S_1 - S_2}{\sqrt{\dfrac{\sigma_1^2}{2n_1} + \dfrac{\sigma_2^2}{2n_2}}}$

(d) none

ANSWERS

1. (c) 2. (c) 3. (b) 4. (c) 5. (b) 6. (c) 7. (c) 8. (a)
9. (b) 10. (b) 11. (b) 12. (c) 13. (c)

8 Test of Hypothesis (Small Samples)

Prerequisites

Before you start reading this unit, you should:

- Be thorough in all the steps of testing procedure
- Be able to calculate mean and variance for a grouped and ungrouped data

Learning Objectives

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

- Understand the concepts of small samples, null and alternative hypothesis, and procedure involved in testing them
- Formulate an appropriate null hypothesis, alternative hypothesis, and all other steps involved in testing procedure
- Differentiate between different tests like when to apply *t*-test or *F*-test or χ^2-test.
- Specify the most appropriate test of hypothesis in a given situation, apply the procedure and make inferences from the result.

INTRODUCTION

In the previous unit, we have discussed different tests of hypothesis for large samples and where the sample size is very large we generally take normal test for testing any statistics. In this unit, we shall look at some of the tests of hypothesis for small samples, that is, samples whose size is less than 30. When the sample size is less than 30, normal tests usually do not hold and hence we go for an appropriate small sample test for testing any statistics.

Some of the tests that are observed in small sample theory are *t*-test, *F*-test, and χ^2-test.

8.1 STUDENT'S *t*-DISTRIBUTION

This distribution is named after a great statistician William Sealy Gosset (1876–1937) and the word 'student' was the pseudonym of him.

If we calculate the means of small samples collected from a large population and then plot the frequency distribution of these means then the resulting sampling distribution would be the student's *t*-distribution. This is almost similar to a normal curve, but a little flatter.

In addition 95 per cent limits will lie farther from the mean in the *t*-distribution than they do in normal distribution. The way in which the sampling distribution of small samples differs from the normal curve is that it changes with the sample size. As the size approaches 30, the *t*-distribution becomes more and more normal.

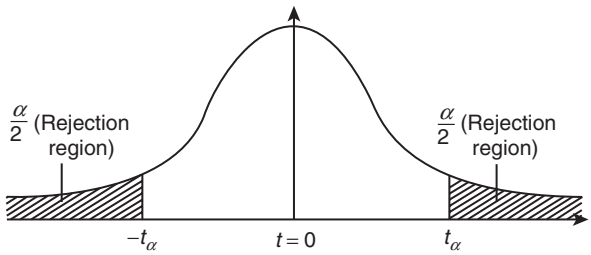
Let x_i, $i = 1, 2, \ldots n$ be a random sample of size n from a normal population with mean μ and variance σ^2. Then student's t is defined by

$$t = \frac{\bar{x} - \mu}{\frac{S}{\sqrt{n}}}$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$.

Here, the sample mean and S^2 is unbiased estimate of the population variance. This follows student's t-distribution with $\gamma = (n - 1)$ degrees of freedom. For student's t-distribution, the number of degrees of freedom is the sample size minus one, that is, $(n - 1)$.

8.2 CRITICAL VALUES OF t



Since t-distribution is symmetric around $t = 0$, the critical values at α level of significance for a one-tailed test (right or left) can be obtained from the table of two-tailed test by looking at the critical value at level of significance 2α.

For example, $t_\gamma(0.05)$ for one-tailed test = $t_\gamma(0.10)$ for two-tailed test.

Hence, the critical value or significant value of t at level of significance α and γ degrees of freedom for two-tailed test are given by,

$p[|t| > t_\gamma(\alpha)] = \alpha$

$p[|t| \le t_\gamma(\alpha)] = 1 - \alpha$

$\therefore\ p[t > t_\gamma(\alpha)] + p[t < -t_\gamma(\alpha)] = \alpha$

$2p[t > t_\gamma(\alpha)] = \alpha$

$\therefore\ p[t > t_\gamma(\alpha)] = \frac{\alpha}{2}$

or

$p[t > t_\gamma(2\alpha)] = \alpha$

8.3 t-TEST FOR SINGLE MEAN

Suppose we want to test

(i) If a random sample of size n has been drawn from a normal population with a specified mean μ_0

or

(ii) If the sample mean differs from the parameter value μ_0 significantly, then under the null hypothesis H_0:

- The sample has been drawn from the population with mean μ

or

- There is no significant difference between sample mean $\bar{x}$ and the population mean μ.

The test statistic used is

$$t = \frac{\bar{x} - \mu_0}{\frac{S}{\sqrt{n}}}$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, $S^2 = \frac{1}{n-1}\sum (x_i - \bar{x})^2$ follows student's *t*-distribution with $(n - 1)$ degrees of freedom.

Conclusion: If the calculated value of *t* namely $|t|$ is greater than tabulated value of *t* for $(n - 1)$ degrees of freedom and α level of significance, null hypothesis is rejected, otherwise it is accepted.

8.4 *t*-TEST FOR DIFFERENCE OF MEANS

Consider two independent samples $x_i (i = 1, 2, \ldots n_1)$ of size n_1 and $y_j (j = 1, 2, \ldots n_2)$ of size n_2, respectively.

Suppose we want to test if the above two independent samples have been drawn from two normal populations with means μ_x and μ_y, respectively.

Under the null hypothesis H_0,

The samples have been drawn from two normal populations with means μ_x and μ_y then the statistic is

$$t = \frac{(\bar{x} - \bar{y}) - (\mu_x - \mu_y)}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

where $\bar{x} = \frac{1}{n_1}\sum_{i=1}^{n_1} x_i$, $\bar{y} = \frac{1}{n_2}\sum_{j=1}^{n_2} y_i$,

$$S^2 = \frac{1}{n_1 + n_2 - 2}\left[\sum_i (x_i - \bar{x})^2 + \sum_j (y_i - \bar{y})^2\right]$$

or

$$S^2 = \frac{1}{n_1 + n_2 - 2}\left[n_1 S_1^2 + n_2 S_2^2\right]$$

$$S_1^2 = \frac{1}{n_1 - 1}\sum_i (x_i - \bar{x})^2$$

$$S_2^2 = \frac{1}{n_2 - 1}\sum_j (y_j - \bar{y})^2$$

follows student's *t*-distribution with $(n_1 + n_2 - 2)$ degrees of freedom.

Caution:

If we want to test if the two independent sample means have been drawn from the populations with same means, that is, under H_0, the samples have been drawn from two populations with $\mu_x = \mu_y$, then the test statistic is $t = \frac{\bar{x} - \bar{y}}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$ follows student's *t*-distribution with $(n_1 + n_2 - 2)$ degrees of freedom.

Worked Out Examples

EXAMPLE 8.1

The heights of 10 adult males selected at random from a given locality had a mean of 158 cm and variance 39.0625 cm. Test at 5% level of significance, the hypothesis that the adult males of the given locality are on the average less than 162.5 cm tall.

Solution:

(i) **Null hypothesis:** The average height of adult males is 162.5 cm.

$$H_0: \mu = 162.5 \text{ cm}$$

(ii) **Alternative hypothesis:** $H_1: \mu < 162.5$ (left-tailed test)

(iii) **Level of significance:** $\alpha = 5\%$ level. The critical value of t for $n - 1 = 10 - 1 = 9$ degrees of freedom and 5% level of significance is $t = 1.83$

(iv) **Test statistic:** Under H_0, the test statistic is

$$t = \frac{\bar{x} - \mu}{\frac{S}{\sqrt{n}}}$$

$$\text{Given } \bar{x} = 158 \text{ cm}, \ \mu = 162.5 \text{ cm}$$

$$S^2 = 39.0625, \ n = 10$$

$$t = \frac{158 - 162.5}{\sqrt{\frac{39.0625}{10}}} = \frac{-4.5}{1.9764}$$

$$= -2.2768$$

(v) **Conclusion:** Since the calculated value of t which $|t| = 2.2768$ is greater than critical value of t which is $t_\alpha = 1.83$ for 9 degrees of freedom, the null hypothesis is rejected and hence alternative hypothesis is accepted. Hence the average height of adult males in the given locality is less than 162.5 cm tall.

EXAMPLE 8.2

Certain pesticide is packed into bags by a machine. A random sample of 10 bags is drawn and their contents are found to weigh in kgs as follows:

50, 49, 44, 52, 45, 48, 46, 45, 49, 45

Test if the average packing can be taken as 50 kg.

Solution: To find $\bar{x}$ and s of the above sample,

$$\bar{x} = \frac{\sum x_i}{n} = \frac{1}{10}[50 + 49 + 44 + 52 + 45 + 48 + 46 + 45 + 49 + 45]$$

$$\bar{x} = \frac{473}{10} = 47.3$$

$$S^2 = \frac{1}{n}\sum (x_i - \bar{x})^2 = \frac{1}{(10-1)}[(50 - 47.3)^2 + (49 - 47.3)^2 + (46 - 47.3)^2 + (52 - 47.3)^2$$
$$(45 - 47.3)^2 + (48 - 47.3)^2 + (46 - 47.3)^2 + (49 - 47.3)^2$$
$$(45 - 47.3)^2 + (45 - 47.3)^2]$$

$$= \frac{1}{(10-1)}[7.29 + 2.89 + 10.89 + 22.09 + 5.29 + 0.49 + 1.69 + 2.89 + 5.29 + 5.29]$$

$$= \frac{64.1}{9} = 7.122$$

$$S = 2.668$$

(i) **Null hypothesis:** H_0: $\mu = 50$ kg the average packing of bags is 50 kg.

(ii) **Alternative hypothesis:** H_1: $\mu \neq 50$ kg (two-tailed test)

(iii) **Level of significance:** $\alpha = 5\%$. The critical value of t for $\gamma = n - 1 = 10 - 1 = 9$ degrees of freedom is t.

$$t_{0.1} = 1.383 (t_{0.05} \text{ (for single tail)} = t_{0.1} \text{ for two-tailed test})$$

(iv) **Test statistic:** Under H_0, $t = \dfrac{\bar{x} - \mu}{\dfrac{8}{\sqrt{n}}} = \dfrac{47.3 - 50}{\dfrac{2.668}{\sqrt{10}}}$

$$t = \frac{-2.7}{0.8436} = -3.2002$$

(v) **Conclusion:** Since the calculated value of t, $|t| = 3.2002$ is greater than the critical value of $t = 1.383$ for 9 degrees of freedom, null hypothesis is rejected and alternative hypothesis is accepted at 5% level of significance. Hence, the average packing of bags ≠ 50 kg.

EXAMPLE 8.3

The life time of electric bulbs for a random sample of 10 from a large consignment gave the following data:

Life in thousand hours: 4.2, 4.6, 4.1, 3.9, 5.2, 3.8, 3.9, 4.3, 4.4, 5.6.

Can we accept the hypothesis that the average life time of bulbs is 4000 hours?

Solution:

$$\bar{x} = \frac{1}{n} \sum x_i = \frac{1}{10}[44] = 4.4$$

$$S^2 = \frac{1}{10-1}[(4.2-4.4)^2 + (4.6-4.4)^2 + (4.1-4.4)^2 + (3.9-4.4)^2 + (5.2-4.4)^2 +$$

$$(3.8-4.4)^2 + (3.9-4.4)^2 + (4.3-4.4)^2 + (4.4-4.4)^2 + (5.6-4.4)^2]$$

$$= \frac{1}{9}[0.04 + 0.04 + 0.09 + 0.25 + 0.64 + 0.36 + 0.025 + 0.01 + 1.44]$$

$$S^2 = \frac{2.895}{9} = 0.3216$$

$$\therefore S = 0.567$$

(i) **Null hypothesis:** H_0: $\mu = 4$, that is, the average life time hours of bulbs is 4000 hours.

(ii) **Alternative hypothesis:** H_1: $\mu \neq 4$ (two-tailed test)

(iii) **Level of significance:** $\alpha = 5\%$. The critical value of t for $\gamma = 10 - 1 = 9$ degrees of freedom is $t_{0.1} = 1.383$.

(iv) **Test statistic:** Under H_0, the test statistic is

$$t = \frac{\bar{x} - \mu}{\dfrac{S}{\sqrt{n}}} = \frac{4.4 - 4}{\dfrac{0.567}{\sqrt{10}}} = \frac{0.4}{0.1793}$$

$$t = 2.2308$$

(v) **Conclusion:** Since the calculated value of t *is* $|t| = 2.2308$ which is greater than critical value of $t = 1.383$, null hypothesis is rejected and alternative hypothesis is accepted.

Hence the average life time of electric bulbs is not 4,000 hours.

EXAMPLE 8.4

An automobile tyre manufacturer claims that the average life of a particular grade of tyre is more than 20,000 km when used under normal driving conditions. A random sample of 16 tyres was tested and a mean of 22,000 and standard deviation of 5000 km was computed. Assuming the lives of the tyres in km to be approximately normally distributed, decide whether the manufacturer's product is as good as claimed.

Solution:

(i) **Null hypothesis:** H_0: $\mu = 20{,}000$ kms

The average life of a particular grade of tyre is 20,000 km when used under normal driving conditions.

(ii) **Alternative hypothesis:** H_1: $\mu > 20{,}000$ (right-tailed test)

(iii) **Level of significance:** $\alpha = 5\%$. The critical value of t for $\gamma = 16 - 1 = 15$ degree of freedom $t_{0.05} = 1.753$

(iv) **Test statistic:** Under H_0, the test statistic is

$$t = \frac{\bar{x} - \mu}{\frac{S}{\sqrt{n}}} = \frac{22000 - 20000}{\frac{5000}{\sqrt{16}}} = \frac{2000}{\frac{5000}{4}}$$

$$= 1.6$$

(v) **Conclusion:** Since the calculated value of $t = 1.6$ is less than tabulated value (critical value) of t, $t_{0.05} = 1.753$, null hypothesis is accepted.

Hence, the average life of a particular grade of tyre is 20,000 km when used under normal driving conditions.

EXAMPLE 8.5

Samples of two types of electric light bulbs were tested for length of life and the following data were obtained:

	Type I	Type II
Sample number	$n_1 = 8$	$n_2 = 7$
Sample means	$\bar{x}_1 = 1,2,3,4$ hours	$\bar{x}_2 = 1,036$ hours
Sample SD's	$S_1 = 36$ hours	$S_2 = 40$ hours

Is the difference in the means significant to warrant that Type I is superior to Type II regarding the length of life?

Solution:

(i) **Null hypothesis:** H_0: There is no significant difference in the means of Type I and Type II regarding length of life, that is, H_0: $\mu_1 = \mu_2$.

(ii) **Alternative hypothesis:** H_1: Type I is superior to Type II regarding length of life.

H_1: $\mu_1 > \mu_2$ (right-tailed test)

(iii) **Level of significance:** $\alpha = 5\%$. The critical value of t for $n_1 + n_2 - 2 = 8 + 7 - 2 = 13$ degrees of freedom and 5% level of significance is

$$t_\alpha = 1.771$$

(iv) **Test statistic:** Under H_0, the test statistic is

$$t = \frac{\bar{x} - \bar{y}}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$S^2 = \frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2 - 2} = \frac{8(36)^2 + 7(40)^2}{8 + 7 - 2} = \frac{21568}{13}$$

$$S = 101.887$$

$$t = \frac{1234 - 1036}{101.887\sqrt{\frac{1}{8} + \frac{1}{7}}} = \frac{198}{101.887\sqrt{(0.125 + 0.1428)}} = \frac{198}{52.7315}$$

$$t = 3.7548$$

(v) **Conclusion:** Since the calculated value of $t = 3.7548$ is greater than critical value of $t_\alpha = 1.771$, null hypothesis is rejected and alternative hypothesis is accepted. Hence Type I electric bulbs are superior to Type II regarding length of life.

EXAMPLE 8.6

Of the two salesmen, *X* claims that he has made more sales than *Y*. For the accounts examined, which were comparable for the two salesmen, the following results were obtained:

	X	*Y*
Number of sales	10	17
Average size of sales	₹6,200	₹5,600
Standard deviation of sales	₹690	₹600

Do these two "average size of sales" differ significantly?

Solution:

(i) **Null hypothesis:** H_0: $\mu_1 = \mu_2$, the average sizes of sales of the two salesmen do not differ significantly.

(ii) **Alternative hypothesis:** H_1: $\mu_1 \neq \mu_2$ (two-tailed test)

(iii) **Level of significance:** $\alpha = 5\%$. The critical value of t for $\gamma = n_1 + n_2 - 2 = 10 + 17 - 2 = 25$ degrees of freedom and 5% level of significance is $t_{0.15} = 1.316$.

(iv) **Test statistic:** Under null hypothesis, the test statistic is

$$t = \frac{\bar{x} - \bar{y}}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$\text{where } S^2 = \frac{n_1S_1^2 + n_2S_2^2}{n_1 + n_2 - 2} = \frac{10(690)^2 + 17(600)^2}{10 + 17 - 2} = 435240$$

$$S = 659.725, \quad \bar{x} = 6200, \quad \bar{y} = 5600$$

$$t = \frac{6200 - 5600}{659.725\sqrt{\frac{1}{10} + \frac{1}{17}}} = \frac{600}{659.725\sqrt{0.1 + 0.0588}}$$

$$= \frac{600}{659.725(0.3985)} = \frac{600}{262.918} = 2.282$$

(v) **Conclusion:** Since the calculated value of t is $t_{0.1} = 2.282$ is greater than the critical value of t, $t_{0.1} = 1.316$, we reject null hypothesis at 5% level of significance and accept alternative hypothesis.

Hence the average sizes of sales of the two salesmen are significantly different.

EXAMPLE 8.7

Two types of batteries X and Y are tested for their length of life and the following result are obtained:

Battery	Sample size	Mean (hours)	Variance (hours)
A	10	1000	100
B	12	2000	121

Is there a significant difference in the two means?

Solution:

(i) **Null hypothesis:** H_0: $\mu_1 = \mu_2$. There is no significant difference between the average length of lives of the two types of batteries X and Y.

(ii) **Alternative hypothesis:** H_1: $\mu_1 \neq \mu_2$ (two-tailed test)

(iii) **Level of significance:** $\alpha = 5\%$. The critical value of t for $\gamma = n_1 + n_2 - 2 = 10 + 12 - 2 = 20$ degrees of freedom and 5% level of significance is $t_{0.05}$ (for one-tailed) $= t_{0.1}$ (two-tailed) $= 1.325$.

(iv) **Test statistic:** Under null hypothesis, the test statistic,

$$t = \frac{\bar{x} - \bar{y}}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$\text{where } S^2 = \frac{n_1S_1^2 + n_2S_2^2}{n_1 + n_2 - 2}$$

$$S^2 = \frac{10(100) + 12(121)}{10 + 12 - 2} = \frac{2452}{20} = 122.6$$

$$S = 11.0724, \quad \bar{x} = 1000, \quad \bar{y} = 2000$$

$$t = \frac{1000 - 2000}{11.0724\sqrt{\frac{1}{10} + \frac{1}{12}}} = \frac{-2000}{11.0724\sqrt{0.1 + 0.833}}$$

$$= \frac{-1000}{11.0724(0.1833)} = \frac{-1000}{2.02994} = -492.6$$

(v) **Conclusion:** Since $|t_{cal}| = 492.6$ is very much greater than the critical value of t, we reject null hypothesis and accept alternative hypothesis. Hence, the average lengths of lives are significantly different for the two types of batteries X and Y.

EXAMPLE 8.8

In a test given to two groups of students, the marks obtained are as follows:

First Group	18	20	36	50	49	36	34	49	41
Second Group	29	28	26	35	30	44	46		

Test if the average marks secured by two groups of students differ significantly.

Solution:

(i) **Null hypothesis:** H_0: $\mu_1 = \mu_2$. There is no significant difference between average marks secured by two groups of students.

(ii) **Alternative hypothesis:** H_1: $\mu_1 \neq \mu_2$ (two-tailed test)

(iii) **Level of significance:** $\alpha = 5\%$. The critical values of t for 5% level of significance and $\gamma = n_1 + n_2 - 2 = 9 + 7 - 2$. $\lambda = 14$ degrees of freedom is $t_{0.1} = 1.345$

(iv) **Test statistic:** Under H_0, the test statistic is

$$t = \frac{\bar{x} - \bar{y}}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$\text{where } S = \frac{\sum (x_i - \bar{x})^2 + \sum (y_j - \bar{y})^2}{n_1 + n_2 - 2}$$

$$\bar{x} = \frac{\sum x_i}{n} = \frac{1}{9}[18 + 20 + 36 + 50 + 49 + 36 + 34 + 49 + 41]$$

$$= \frac{1}{9}[333]$$

$$= 37$$

$$\sum (x_i - \bar{x})^2 = (18-37)^2 + (20-37)^2 + (36-37)^2 + (50-37)^2 + (49-37)^2 + (36-37)^2 + (34-37)^2 + (49-37)^2 + (41-37)^2$$

$$= 361 + 289 + 1 + 169 + 144 + 1 + 9 + 144 + 16$$

$$= 1134$$

$$\bar{y} = \frac{\sum y_j}{n} = \frac{1}{7}[29 + 28 + 26 + 35 + 30 + 44 + 46]$$

$$= \frac{1}{7}(238)$$

$$= 34$$

$$\sum (y_j - \bar{y})^2 = (29-34)^2 + (28-34)^2 + (26-34)^2 + (35-34)^2 + (30-34)^2 + (44-34)^2 + (46-34)^2$$

$$= 25 + 36 + 64 + 1 + 16 + 100 + 144$$

$$= 386$$

$$S^2 = \frac{\sum(x_i - \bar{x})^2 + \sum(y_j - \bar{y})^2}{n_1 + n_2 - 2} = \frac{1134 + 384}{9 + 7 - 2} = 108.428$$

$$S = 10.413, \quad \bar{x} = 37, \quad \bar{y} = 34$$

$$t = \frac{\bar{x} - \bar{y}}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{37 - 34}{10.413\sqrt{\frac{1}{9} + \frac{1}{7}}} = \frac{3}{10.413\sqrt{(0.11 + 0.1428)}}$$

$$= \frac{3}{10.413(0.5038)} = \frac{3}{5.246} = 0.5718$$

(v) **Conclusion:** Since the calculated vale of $t = 0.5718$ is less than critical value of $t = 1.345$, we accept null hypothesis. Hence, there is no significant difference between the average marks secured by two groups of students.

EXAMPLE 8.9

For a random sample of 10 persons fed on diet *A* the increased weight in pounds in a certain period were 10, 6, 16, 17, 13, 12, 8, 14, 15, and 9. For another random sample of 12 persons fed on diet *B*, the increase in the same period were 7, 13, 22, 15, 12, 14, 18, 8, 21, 23, 10, and 17. Test whether the diets *A* and *B* differ significantly as regards their effect on increase in weight.

Solution:

(i) **Null hypothesis:** H_0: $\mu_1 = \mu_2$, there is no significant difference in the diets *A* and *B* with regard to their effect on increase in weight.

(ii) **Alternative hypothesis:** H_1: $\mu_1 \neq \mu_2$ (two-tailed test)

(iii) **Level of significance:** $\alpha = 5\%$. The critical value of t for $\gamma = n_1 + n_2 - 2 = 10 + 12 - 2 = 20$ degree of freedom is $t_{0.1} = 1.325$

(iv) **Test statistic:** Under H_0, the test statistic is

$$t = \frac{\bar{x} - \bar{y}}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \text{ where } S = \frac{\sum(x_i - \bar{x})^2 + \sum(y_j - \bar{y})^2}{n_1 + n_2 - 2}$$

$$\bar{x} = \frac{1}{n}\sum x_i = \frac{1}{10}[10 + 6 + 16 + 17 + 13 + 12 + 8 + 14 + 15 + 9] = 12$$

$$\bar{y} = \frac{1}{n}\sum y_j = \frac{1}{12}[7 + 13 + 22 + 15 + 12 + 14 + 18 + 8 + 21 + 23 + 10 + 17] = 15$$

$$\sum(x_i - \bar{x})^2 = 4 + 36 + 16 + 25 + 1 + 0 + 16 + 4 + 9 + 9 = 102$$

$$\sum(y_j - \bar{y})^2 = 64 + 4 + 49 + 0 + 9 + 1 + 9 + 49 + 36 + 64 + 25 + 4 = 314$$

$$S^2 = \frac{120 + 314}{10 + 12 - 2} = 21.7, \quad S = 4.658$$

$$t = \frac{12 - 15}{4.658\sqrt{\frac{1}{10} + \frac{1}{12}}} = \frac{-3}{4.658(0.428)} = \frac{-3}{1.994} = -1.5045$$

(v) **Conclusions:** Since $|t| = 1.5045$ is slightly greater than critical value $t = 1.325$, we reject null hypothesis and accept alternative hypothesis. Hence, the diets A and B are significantly different with regard to their effect on increase in weight.

8.5 PAIRED *t*-TEST FOR DIFFERENCE OF MEANS

In the earlier t-test, the two samples were independent of each other. Let us now take a particular situation where,

(i) The size of samples are equal, that is, $n_1 = n_2 = n$

(ii) The sample observations, that is, $(x_1, x_2, \ldots x_{n1})$ and $(y_1, y_2, \ldots y_{n_2})$ are not completely independent, but they are dependent in pairs, that is, the sample observations are paired together (x_i, y_i) $i = 1,2 \ldots n$ corresponding to the same i^{th} sample unit.

The problem is to test if the sample means differ significantly or not.
Let $d_i = x_i - y_i$, $i = 1,2, \ldots n$ denote the difference in the observations for the i^{th} unit.
Under the null hypothesis, H_0: the increments are just by chance and not due to advertisement campaign, that is, $H_0: \mu_1 = \mu_2$, the test statistic used is

$$t = \frac{\bar{d}}{\frac{S}{\sqrt{n}}}$$

$$\text{where } \bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i, \qquad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(d_i - \bar{d})^2$$

follows student's t-distribution, with $(n - 1)$ degrees of freedom.

Worked Out Examples

EXAMPLE 8.10

Ten accountants were given intensive coaching and four tests were conducted in a month. The scores of tests 1 and 4 are as follows:

Number of accountant	1	2	3	4	5	6	7	8	9	10
Marks in 1st test	50	42	51	42	60	41	70	55	62	38
Marks in 4th test	62	40	61	52	68	51	64	63	72	50

Does the score from test 1 to test 4 shows an improvement at 5% level of significance?

Solution:

(i) **Null hypothesis:** H_0: There is no significant difference in the test scores from test 1 to test 4.

$$H_0: \mu_1 = \mu_2$$

(ii) **Alternative hypothesis:** $H_1: \mu_1 \neq \mu_2$ (two-tailed test)

(iii) **Level of significance:** $\alpha = 5\%$. The critical value of t for $\gamma = n - 1 = 10 - 1 = 9$ degree of freedom for 5% level of significance, $t_{0.1} = 1.372$

(iv) **Test statistic:** Under H_0, the test statistic, that is,

$$t = \frac{\bar{d}}{\frac{S}{\sqrt{n}}}$$

where $\bar{d} = \frac{1}{n}\sum d_i$, $S^2 = \frac{1}{n-1}\sum (d_i - \bar{d})^2$

x_i	y_i	$d_i = x_i - y_i$	$(d_i - \bar{d})^2$
50	62	−12	23.04
42	40	2	84.64
51	61	−10	7.84
42	52	−10	7.84
60	68	−8	0.64
41	51	−10	7.84
70	64	6	1.44
55	63	−8	0.64
62	72	−10	7.84
38	50	−12	23.04
Total		−72	164.8

$$S^2 = \frac{1}{n-1}(164.8)$$

$$= \frac{1}{9}(164.8)$$

$$= 18.311$$

$$S = 4.279$$

$$t = \frac{-7.2}{\frac{4.279}{\sqrt{10}}} = \frac{-7.2}{1.353}$$

$$t = 5.3210$$

$$\bar{d} = \frac{1}{10}\sum(-72) = -7.2$$

(v) **Conclusion:** Since the calculated value of t, $|t| = 5.3210$ is greater than the critical value of $t = 1.372$, we reject null hypothesis and accept alternative hypothesis. Hence, there is significant difference in the test scores of tests 1 and 4.

EXAMPLE 8.11

A drug is given to 10 patients and the increments in their blood pressure were recorded to be 3, −6, −2, 4, −3, 4, 6, 0, 0, and 2. Is it reasonable to believe that the drug has no effect on change of blood pressure? Test at 1% end.

Solution:

(i) **Null hypothesis:** H_0: The increments in the blood pressure of 10 patients are by chance, that is, the drug has no effect on change of blood pressure.

(ii) **Alternative hypothesis:** The drug has an effect on change of blood pressure.

(iii) **Level of significance:** $\alpha = 1\%$. The critical value of t for $\gamma = n - 1 = 10 - 1 = 9$ degrees of freedom is $t_{0.2} = 0.883$

(iv) **Test statistic:** Under H_0, the test statistic is

$$t = \frac{\bar{d}}{\frac{S}{\sqrt{n}}} \text{ where } S = \frac{1}{n-1}\sum (d_i - \bar{d})^2$$

$$\bar{d} = \frac{1}{n}\sum d_i = \frac{8}{10} = 0.8$$

$$S = \frac{1}{9}[(3-0.8)^2 + (-6-0.8)^2 + (-2-0.8)^2 + (y-0.8)^2 + (-3-0.8)^2$$
$$+ (4-0.8)^2 + (6-0.8)^2 + (2-0.8)^2]$$
$$= \frac{1}{9}[4.84 + 46.24 + 7.84 + 10.24 + 14.44 + 10.24 + 27.04 + 1.44] = 13.59,$$

$$S = 3.6866$$

$$t = \frac{0.8}{\left(\frac{3.6866}{\sqrt{10}}\right)} = \frac{0.8}{1.1658} = 0.6862$$

(v) **Conclusion:** Since the calculated value of $t = 0.6862$ is less than critical value of $t = 0.883$, we accept null hypothesis. Hence, the drug has no effect on change of blood pressure.

EXAMPLE 8.12

The following table gives the additional hours of sleep gained by 10 patients in an experiment to test the effect of a drug. Do these data give evidence that the drug produces additional hours of sleep?

Patients	1	2	3	4	5	6	7	8	9	10
Hours gained	0.7	0.1	0.2	1.2	0.31	0.4	3.7	0.8	3.8	2.0

Solution:

(i) **Null hypothesis:** H_0: The drug does not provide additional hours of sleep.

(ii) **Alternative hypothesis:** H_1: The drug provides additional hours of sleep.

(iii) **Level of significance:** $\alpha = 5\%$ the critical value of t for $\gamma = n - 1 = 10 - 1 = 9$ degrees of freedom for 5% level of significance is $t_{0.1} = 1.383$.

(iv) **Test statistic:** Under H_0, the test statistic is

$$t = \frac{\bar{d}}{\frac{S}{\sqrt{n}}}, \text{ where } S = \frac{1}{n-1}\sum (d_i - \bar{d})^2$$

$$\bar{d} = \frac{1}{n}\sum d_i = \frac{1}{10}[0.7 + 0.1 + 0.2 + 1.2 + 0.31 + 0.4 + 3.7 + 0.8 + 3.8 + 2.0] = \frac{13.21}{10}$$

$$= 1.321$$

$$S^2 = \frac{1}{10-1}[(0.7-1.321)^2 + (0.1-1.321)^2 + (0.2-1.321)^2 + (1.2-1.321)^2 + (0.31-1.321)^2 + (0.4-1.321)^2 + (3.7-1.321)^2 + (0.8-1.321)^2 + (3.8-1.321)^2 + (2-1.321)^2]$$

$$= \frac{1}{9}[0.3856 + 1.4908 + 1.2566 + 0.0146 + 1.0221 + 0.8482 + 5.6596 + 0.2714 + 6.1454 + 0.46104]$$

$$= \frac{1}{9}[17.555] = 1.9505, \quad S = 1.3966$$

$$t = \frac{1.321}{\left(\frac{1.3966}{\sqrt{10}}\right)} = \frac{1.321}{0.44164}$$

$$t = 2.99112$$

(v) **Conclusion:** Since the calculated value of $t = 2.99112$ is greater than the critical value of $t = 1.383$, we reject null hypothesis and accept alternative hypothesis. Hence, the drug provides additional hours of sleep.

8.6 SNEDECOR's *F*-DISTRIBUTION

This is named in honour of a great statistician R. A. Fisher.

F-test for Equality of Population Variances

The various test of significances discussed earlier are not suitable for test of significance of two or more sample estimates of population variance. It is also observed that the means of samples drawn randomly from a normal population are normally distributed, whereas the variances of random samples drawn from such a population are not normally distributed, but are skewed positively.

Suppose we want to test:

(i) Whether two independent samples x_i (i = 1, 2, … n_1) and y_j(j = 1, 2, … n_2) have been drawn from the normal population with the same variance σ^2

or

(ii) Whether the two independent estimates of the population variance are homogeneous or not.

Under the null hypothesis, H_0:

(i) $\sigma_x^2 = \sigma_y^2 = \sigma^2$, that is, the population variances are equal

or

(ii) The two independent estimates of population variances are homogenous.

Then the F-statistic is given by,

$$F = \frac{S_x^2}{S_y^2}$$

where $S_x^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(x_i - \bar{x})^2$, $S_y^2 = \frac{1}{n_2 - 1}\sum_{j=1}^{n_2}(y_j - \bar{y})^2$ are two unbiased estimates of common population variance σ^2.

This follows Snedecor's F-distribution with $(n_1 - 1, n_2 - 1) = (\gamma_1, \gamma_2)$ degrees of freedom.

Caution:

- In the calculation of F, the greater of the two variances is taken in the numerator and n_1 corresponds to the greater variance.
- Critical values of F-distribution:

The tables given for F-distribution are the critical values for the right-tailed test.

The critical value $F_\alpha(\gamma_1, \gamma_2)$ at level of significance α and (γ_1, γ_2) degrees of freedom is determined by $P[F > F_\alpha(\gamma_1, \gamma_2)] = \alpha$

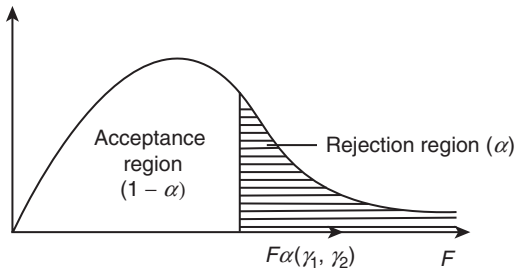
- In addition, we have

$$F\alpha(\gamma_1, \gamma_2) = \frac{1}{F_{1-\alpha}(\gamma_2, \gamma_1)}$$

- The critical values of F for left-tailed test
- H_0: $\sigma_1^2 = \sigma_2^2$ against $H_1: \sigma_1^2 < \sigma_2^2$ are given by

$$F < Fn_1 - 1, n_2 - 1(1 - \alpha)$$

- The critical values of F for the two-tailed test H_0: $\sigma_1^2 \neq \sigma_2^2$ against H_1: $\sigma_1^2 \neq \sigma_2^2$ are given by,



Critical values of F–distribution

$$F > F_{n_1 - 1, n_2 - 1}\left(\frac{\alpha}{2}\right) \text{ and } F < F_{n_1 - 1, n_2 - 1}\left(1 - \frac{\alpha}{2}\right)$$

- This test is also known as variance-ratio test.
- If the computed value of F exceeds the critical value of F for $(n_1 - 1, n_2 - 1)$ degrees of freedom at α% level of significance, null hypothesis is rejected and alternative hypothesis is accepted, otherwise, null hypothesis is accepted.

Worked Out Examples

EXAMPLE 8.13

Two random samples were drawn from two normal populations and their values are as follows:

A	66	67	75	76	82	84	88	90	92		
B	64	66	74	78	82	85	87	92	93	95	97

Test whether the two populations have the same variance at 10% level of significance.

Solution:

(i) **Null hypothesis:** H_0: $\sigma_x^2 = \sigma_y^2$
The two populations have same variance at 5% level.

(ii) **Alternative hypothesis:** H_1: $\sigma_x^2 \neq \sigma_y^2$ (two-tailed test)

(iii) **Level of significance:** $\alpha = 10\%$. The critical value of F for $(\gamma_1, \gamma_2) = (n_1 - 1, n_2 - 1) = (9 - 1, 11 - 1) = (8, 10)$ degrees of freedom is $F > 3.07$.

(iv) **Test statistic:** Under H_0, the test statistic is

$$F = \frac{S_x^2}{S_y^2}$$

where $S_x^2 = \frac{1}{n_1 - 1}\sum (x_i - \bar{x})^2, \quad S_y^2 = \frac{1}{n_2 - 1}\sum (y_j - \bar{y})^2$

$$\bar{x} = \frac{1}{n_1}\sum x_i = \frac{720}{9} = 80, \quad \bar{y} = \frac{1}{n_2}\sum y_j = \frac{913}{11} = 83$$

$$S_x^2 = \frac{1}{9-1}[722] = 90.25, \quad S_y^2 = \frac{1}{11-1}[1298] = 129.8$$

X_i	Y_j	$(X_i - \bar{X})^2$	$(Y_j - \bar{Y})^2$
66	64	196	361
67	66	169	289
75	74	25	81
76	78	16	25
82	82	4	1
84	85	4	4
88	87	64	16
90	92	100	81
92	93	144	100
	95		144
	97		196
720	913	722	1298

$$S_x^2 = 9.5$$

$$S_y^2 = 11.393$$

$$F = \frac{S_y^2}{S_x^2} = \frac{129.8}{90.25}$$

$$= 1.4382$$

Critical value of F:

$$P(F_{8,\ 10} \geq 3.07) = 0.95$$

$$P(F_{10,\ 8} \geq 3.35) = 0.05$$

$$\Rightarrow P(F_{8,\ 10} \geq 3.07) = 0.95$$

$$\text{(i.e.,)} P\left(\frac{1}{F_{8,\ 10}} \leq \frac{1}{3.07}\right) = 0.05$$

$$P(F_{10,\ 8} \leq 0.326) = 0.05$$

$$P(F_{10,\ 8} \geq 0.326) = 0.95$$

(v) **Conclusion:** Since the calculated value of $F = 1.4882$ lies between $F > 3.35$ and $F < 0.326$ and hence null hypothesis is accepted.

Hence, the two populations have the same variance at 10% level of significance.

EXAMPLE 8.14

Can the following samples be regarded as coming from the same normal populations?

Sample	Size	Sample mean	Sum of squares of durations from the mean
1	10	12	120
2	12	15	314

Solution: To test if the two independent samples have been drawn from the same normal population, we have to test the following:

(i) The equality of population means

(ii) Equality of population variances

The t-test is used for testing equality of means and F-test is used for testing equality of variances.

(i) **Null hypothesis:** The samples can be regarded as drawn from normal population.

$$H_0: \mu_1 = \mu_2 \text{ and } \sigma_1^2 = \sigma_2^2$$

(ii) **Alternative hypothesis:** $H_1: \mu_1 \neq \mu_2$ and $\sigma_1^2 \neq \sigma_2^2$

(iii) **Level of significance:** $\alpha = 5\%$. The critical values of t for $\gamma = n_1 + n_2 - 2 = 10 + 12 - 2 = 20$ degrees of freedom are $t_{0.05}$ (for one-tailed) = $t_{0.1}$ (for two-tailed) = 1.325.
The critical value of F is calculated later.

(iv) **Test statistic:** Under H_0, the test statistic is

$$t = \frac{\bar{x} - \bar{y}}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$\text{where } S^2 = \frac{\sum (x_i - \bar{x})^2 + \sum (y_j - \bar{y})^2}{n_1 + n_2 - 2}$$

$$= \frac{120 + 314}{10 + 12 - 2} = 21.7$$

$$S = 4.6583,\ \bar{x} = 12,\ \bar{y} = 15,\ n_1 = 10,\ n_2 = 12$$

$$t = \frac{12 - 15}{4.6583\sqrt{\frac{1}{10} + \frac{1}{12}}} = \frac{-3}{1.9945} = -1.5041$$

Under H_0, the test statistic is

$$F = \frac{S_x^2}{S_y^2},\ S_x^2 = \frac{1}{n_1 - 1}\sum (x_i - \bar{x})^2 = \frac{1}{10 - 1}(120)$$

$$= 13.33$$

$$S_y^2 = \frac{1}{n_2 - 1}\sum (y_j - \bar{y})^2 = \frac{1}{12 - 1}(314) = 28.545$$

$$\therefore F = \frac{S_y^2}{S_x^2} = \frac{28.545}{13.33} = 2.140909$$

(v) **Conclusion:** Since the calculated value of $|t| = 1.5041$ is greater than the critical values $t = 1.325$, we reject null hypothesis and accept alternative hypothesis.

The critical value of F is $F > F_{12-1}$, $10 - 1\left(\frac{\alpha}{2}\right) = 0.05$ and $F < F_{11,9}\left(1 - \frac{\alpha}{2}\right) = F_{11,9}(0.95)$

$$P(F_{11,9} > 3.07) = 0.05 \Rightarrow F_{11,9}(0.05) = 3.07$$

$$P(F_{9,11} > 2.95) = 0.05 \Rightarrow P\left(\frac{1}{F_{9,11}} \le \frac{1}{2.95}\right) = 0.05$$

$$P(F_{11,9} \le 0.338) = 0.05$$

Since the calculated value of $F = 2.14$ lies between $F > 3.07$ and $F < 0.338$, null hypothesis is accepted. Hence, samples are drawn from normal population with equal variance but unequal means.

EXAMPLE 8.15

The following data presents the yields in quintals of common 10 subdivisions of equal area of two agricultural plots:

Plot 1	6.2	5.7	6.5	6.0	6.3	5.8	5.7	6.0	6.0	5.8
Plot 2	5.6	5.9	5.6	5.7	5.8	5.7	6.0	5.5	5.7	5.5

Test whether the two samples taken from the two random populations have the same variance.

Solution:

(i) **Null hypothesis:** The two random populations have the same variance.

$$H_0 : \sigma_1^2 = \sigma_2^2$$

Plot 1 x_i	Plot 2 y_j	$(x_i - \bar{x})^2$	$(y_j - \bar{y})^2$
6.2	5.6	0.04	0.01
5.7	5.9	0.09	0.04
6.5	5.6	0.25	0.01
6.0	5.7	0	0
6.3	5.8	0.09	0.01
5.8	5.7	0.04	0
5.7	6.0	0.09	0.09
6.0	5.5	0	0.04
6.0	5.7	0	0
5.8	5.5	0.04	0.04
60	57	0.64	0.24

$$S_1^2 = \frac{1}{n-1}\Sigma(x_i - \bar{x})^2$$

$$= \frac{0.64}{9} = 0.071$$

$$S_2^2 = \frac{1}{n-1}\Sigma(y_j - \bar{y})^2$$

$$= \frac{0.24}{9} = 0.026$$

$$\bar{x} = \frac{\Sigma x_i}{n_1} = \frac{60}{10} = 6, \quad \bar{y} = \frac{\Sigma y_j}{n_1} = \frac{57}{10} = 5.7$$

(ii) **Alternative hypothesis:**

$$H_1 : \sigma_1 \neq \sigma_2^2$$

(iii) **Level of significance:** $\alpha = 10\%$. The critical value of F for $(\gamma_1, \gamma_2) = (n_1 - 1, n_2 - 1) = (9, 9)$ degrees of freedom $F > 3.18$ and $F < 0.314$

(iv) **Test statistic:** Under H_0, $F = \frac{S_1^2}{S_2^2} = 2.7307$

(v) **Conclusion:** Since the calculated value of F lies within $F > 3.18$ and $F < 0.314$, we accept null hypothesis. Hence, the two samples have been taken from two normal populations with same variance.

EXAMPLE 8.16

Consider the following measurements of the heat producing capacity of the coal produced by two mines (in millions of calories per ton):

Mine 1	8260	8130	8350	8070	8340	
Mine 2	7950	7890	7900	8140	7920	7840

Can it be concluded that the two population variances are equal?

Solution:

(i) **Null hypothesis:**

$$H_0: \sigma_x^2 = \sigma_y^2$$

(ii) **Alternative hypothesis:** $H_1: \sigma_x^2 \neq \sigma_y^2$

(iii) **Level of significance:** $\alpha = 5\%$ the critical value of F for $(\gamma_1 - 1, \gamma_2 - 1) = (5 - 1, 6 - 1) = (4, 5)$ degrees of freedom = 5.19

Mine 1 x_i	Mine 2 y_j	$(x_i - \bar{x})^2$	$(y_j - \bar{y})^2$
8260	7950	900	100
8130	7890	10000	2500
8350	7900	14400	1600
8070	8140	25600	40000
8340	7920	12100	400
	7840		
$\sum x_i = 41150$	$\sum y_j = 47640$	$\sum (x_i - \bar{x})^2 = 63000$	$\sum (y_j - \bar{y})^2 = 44600$

$$\bar{x} = \frac{\sum x_i}{n_1}$$
$$= 8230$$

$$\bar{y} = \frac{\Sigma y_i}{n_2}$$
$$= 7940$$

$$S_x^2 = \frac{1}{n_1 - 1}\Sigma (x_i - \bar{x})^2$$
$$= \frac{1}{5-1}(63000)$$
$$= 15750$$

$$S_y^2 = \frac{1}{n_2 - 1}\Sigma (y_j - \bar{y})^2$$
$$= \frac{1}{6-1}(44600)$$
$$= 8920$$

(iv) **Test statistic:** Under H_0, the test statistic is

$$F = \frac{S_x^2}{S_y^2} = \frac{15750}{8920} = 1.76$$

(v) **Conclusion:** Since the calculated value of $F = 1.76$ is less than the tabulated value of $F = 5.19$, we accept null hypothesis. Hence, the two population variances of the two mines are equal.

8.7 CHI-SQUARE DISTRIBUTION

The tests of significance that were discussed so far were based on assumption that the samples had been drawn from normal population. In the earlier tests as *Z*-test, *t*-test, and *F*-test, we have to make assumptions on the population parameters and hence such tests are called parametric tests.

Chi-square test is a non-parametric test. It is easy to compute and can be used without making assumptions about the parameters.

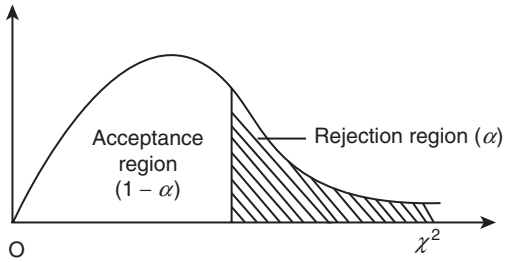
Chi-square Test of "Goodness of Fit"

This test statistic is used in testing a hypothesis that provides a set of theoretical frequencies with which observed frequencies are compared. Hence, the measurement of chi-square gives the degree of discrepancy between the observed frequencies and theoretical frequencies and thus to determine whether the discrepancy so obtained is due to sampling error or due to chance.

Let $O_1, O_2, \ldots, O_n$ be the observed frequencies and let $E_1, E_2, \ldots, E_n$ represent their corresponding expected frequencies, then the chi-square statistic is given by, $\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$ which follows χ^2-distribution with $\gamma = n - 1$ degrees of freedom, under the null hypothesis that there is no significant difference between the observed and theoretical frequencies, that is, there is good compatibility between theory and experiment. The formula is due to Karl Pearson.

Caution:

- The chi-square distribution curve is



- The chi-square values increase with increase in the degrees of freedom and a reduction in the degree of freedom.
- If the calculated value of χ^2 exceeds the critical value of χ^2 at α% level of significance, null hypothesis is rejected, otherwise it is accepted.
- Conditions which validate χ^2-test are follows:
  - The total number of observations which make up the sample for this test must be independent of each other.
  - χ^2-test is wholly based on sample date and no assumption is made concerning the population distribution.
  - The expected frequency of any observation must not be less than 5. If it is less than 5 it should be pooled together with the adjacent cell so that the observation is 5 or more than 5.

Worked Out Examples

EXAMPLE 8.17

The following table gives the number of aircraft accidents that occur during various days of the week. Find whether the accidents are uniformly distributed over the week.

Days of week	Sun	Mon	Tue	Wed	Thu	Fri	Sat
Number of accidents	14	16	8	12	11	9	14

Solution:

(i) **Null Hypothesis:** The accidents are uniformly distributed over the week.

(ii) **Alternative Hypothesis:** The accounts are not uniformly distributed over the week.

(iii) **Level of significance:** $\alpha = 5\%$. The critical value of χ^2 for $n - 1 = 7 - 1 = 6$ degrees of freedom is $\chi^2_{0.05} = 12.592$.

$$\text{Average no. of accidents} = \frac{1}{7}[14+16+8+12+11+9+14]$$

$$= \frac{84}{7} = 12$$

Observed frequencies O_i	Expected frequencies e_i	$(O_i - e_i)^2$	$\frac{(O_i - e_i)^2}{e_i}$
14	12	4	0.333
16	12	16	1.333

8	12	16	1.333
12	12	0	0
11	12	1	0.0833
9	12	9	0.75
14	12	4	0.333
Total			4.1653

(iv) **Test statistic:** Under H_0, the test statistic is

$$\chi^2 = \Sigma \frac{(O_i - e_i)^2}{e_i}$$

$$= 4.1653$$

(v) **Conclusion:** Since the calculated value of $\chi^2 = 4.1653$ is less than the tabulated value of χ^2, $\chi^2_{0.05} = 12.592$, we accept null hypothesis.

Hence, the accidents are uniformly distributed over the week.

EXAMPLE 8.18

A set of 5 coins is tossed 3,200 times and the number of heads appearing each time is noted. The results are as follows:

No. of heads	0	1	2	3	4	5
Frequency	80	570	1,100	900	500	50

Test the hypothesis that coins are unbiased.

Solution:

(i) **Null hypothesis:** H_0: The coins are unbiased.

(ii) **Alternative hypothesis:** H_1: The coins are not unbiased.

(iii) **Level of significance:** $\alpha = 0.05$. The critical value of χ^2 for $\gamma = n - 1 = 6 - 1 = 5$ degrees of freedom are $\chi^2_{0.05} = 11.07$.

(iv) **Test statistic:** Under H_0,

$$\chi^2 = \Sigma \frac{(O_i - e_i)^2}{e_i}$$

$$N = \sum_i f_i$$

$$= 3200$$

Probability of head appearing $= \frac{1}{2}$

Expected number of frequencies are $f(x + 1) = NP(x + 1)$

$$P(x+1) = \left(\frac{n-x}{x+1}\right)\frac{p}{q}P(x), \quad x = 0,1,2,\ldots 4$$

$$P(0) = q^n = \left(\frac{1}{2}\right)^5 = \frac{1}{32}$$

$$f(0) = N \cdot P(0) = \frac{1}{32}(3200) = 100$$

$$P(1) = \frac{5-0}{1} P(0) = \frac{5}{32}$$

$$P(2) = \frac{4}{2} P(1) = \frac{10}{32}$$

$$f(1) = 3200\left(\frac{5}{32}\right) = 500$$

$$f(2) = 3200\left(\frac{10}{32}\right) = 1000$$

$$P(3) = \frac{3}{3} P(2) = \frac{10}{32}$$

$$P(4) = \frac{2}{4} P(3) = \frac{10}{32}\left(\frac{1}{2}\right)$$

$$f(3) = 1000$$

$$f(4) = 3200\left(\frac{5}{32}\right) = 500$$

$$P(5) = \frac{1}{5} P(4) = \frac{1}{32}$$

$$f(5) = 3200\left(\frac{1}{32}\right) = 100$$

Observed frequency O_i	Expected frequency e_i	$(O_i - e_i)^2$	$\frac{(O_i - e_i)^2}{e_i}$
80	100	400	4
570	500	900	1.8
1100	1000	10000	10
900	1000	10000	10
500	500	0	0
50	100	2500	25
3200	3200		50.8
Total			508

(v) **Conclusion:** Since calculated value of χ^2 is very much greater than tabulated value of χ^2 for 5% level of significance, reject H_0 and accept H_1. Hence the coins are biased.

EXAMPLE 8.19

The demand for a particular spare part in a factory was found to vary from day to day. In a sample study the following information was obtained:

Days	Mon	Tue	Wed	Thu	Fri	Sat
No. of parts demanded	1,124	1,125	1,110	1,120	1,126	1,115

Test the hypothesis that the number of parts demanded does not depend on the day of the week.

Solution:

(i) **Null hypothesis:** The number of parts demanded does not depend on the day of the week.

The number of parts demanded during six days = 1124 + 1125 + 1110 + 1120 + 1126 + 1115 = 6720

Hence, on an average $\frac{6720}{6} = 1{,}120$ parts are to be demanded each day of the week.

Observed frequencies O_i	Expected frequencies e_i	$(O_i - e_i)^2$	$\frac{(O_i - e_i)^2}{e_i}$
1124	1120	16	0.9428
1125	1120	25	0.0223
1110	1120	100	0.0892
1120	1120	0	0
1126	1120	36	0.0321
1115	1120	25	0.0223
Total			0.1802

(ii) **Alternative hypothesis:** The parts are not to be demanded each day of the week.

(iii) **Level of significance:** $\alpha = 0.05$. The critical value of χ^2 for $\gamma = 6 - 1 = 5$ degrees of freedom and at 5% level of significance is $\chi^2_{0.05} = 11.070$.

(iv) **Test statistic:** Under H_0, the test statistic is

$$\chi^2 = \frac{\sum (O_i - e_i)^2}{e_i}$$
$$= 0.1802$$

(v) **Conclusion:** Since the calculated value of $\chi^2 = 0.1802$ is less than the tabulated value of χ^2, $\chi^2_{\alpha} = 11.070$ we accept null hypothesis. Hence the number of parts demanded does not depend on the day of the week.

EXAMPLE 8.20

A survey of 800 families with 4 children gave the following distribution:

No. of boys	0	1	2	3	4
No. of girls	4	3	2	1	0
No. of families	32	178	290	236	64

Is this result consistent with the hypothesis that the male and female births are equally probable?

Solution:

(i) **Null hypothesis:** The data are consistent with the hypothesis of equal probability for male and female births, that is,

$$p = \text{probability of male birth} = q = \frac{1}{2}$$

Let $P(x)$ = Probability of 'x' male births in a family of 4

$$= 4C_x \left(\frac{1}{2}\right)^2 \left(\frac{1}{2}\right)^{4-x} \qquad \{\text{Since this follows binomial distribution}\}$$

$$= 4C_x \left(\frac{1}{2}\right)^4$$

The frequencies of x male births is given by

$$f(x) = N \cdot P(x) = 800 \cdot 4C_x \left(\frac{1}{2}\right)^4,$$

$$= 50 \times 4C_x, \qquad x = 0, 1, 2, 3, 4$$

The expected frequencies are as follows:

$$r = 0: \quad f(0) = 50 \times 4C_0 = 50$$
$$r = 1: \quad f(1) = 50 \times 4C_1 = 200$$
$$r = 2: \quad f(2) = 50 \times 4C_2 = 50 \times 6 = 300$$
$$r = 3: \quad f(3) = 50 \times 4C_3 = 200$$
$$r = 4: \quad f(4) = 50 \times 4C_4 = 50$$

(ii) **Alternative hypothesis:** The data is not consistent with the hypothesis of equal probability for male and female probability.

(iii) **Level of significance:** $\alpha = 5\%$. The critical value of χ^2 for $\gamma = n - 1 = 5 - 1 = 4$ degrees of freedom = 9.488.

(iv) **Test statistic:** Under null hypothesis

$$\chi^2 = \Sigma \frac{(O_i - e_i)^2}{e_i}$$

No. of male births	Observed frequency O_i	Expected frequency e_i	$(O_i - e_i)^2$	$\frac{(O_i - e_i)^2}{e_i}$
0	32	50	324	6.48
1	178	200	484	2.42
2	290	300	100	0.33
3	236	200	1296	6.48
4	64	50	196	3.92
Total	800	800		19.63

$$\text{Hence, } \chi^2_{cal} = \sum \frac{(O_i - e_i)^2}{e_i}$$
$$= 19.63$$

(v) **Conclusion:** Since the calculated value of χ^2 is very much higher than the tabulated value of χ^2, we reject null hypothesis. Hence male and female births are not equally probable.

EXAMPLE 8.21

200 digits are chosen at random from a set of tables. The frequencies of the digits are as follows:

Digit	0	1	2	3	4	5	6	7	8	9
Frequency	18	19	23	21	16	25	22	20	21	15

Use χ^2 test to assess the correctness of the hypothesis that the digits were distributed in equal numbers in the tables from which they were chosen.

Solution:

(i) **Null hypothesis:** The digits were distributed in equal numbers.
Then the expected frequencies will be $\frac{200}{10} = 20$ as the frequency of 0, 1, 2, … 9 digits.

(ii) **Alternative hypothesis:** The digits were not distributed in equal numbers.

(iii) **Level of significance:** $\alpha = 5\%$. The critical value of χ^2 for $\gamma = n - 1 = 10 - 1 = 9$ and 5% level of significance is $\chi^2_{0.05} = 16.22$.

(iv) **Test statistic:** Under the null hypothesis, the test statistic is

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - e_i)^2}{e_i}$$

Digit	Observed frequencies O_i	Expected frequencies e_i	$(O_i - e_i)^2$	$\frac{(O_i - e_i)^2}{e_i}$
0	18	20	4	0.20
1	19	20	1	0.05
2	23	20	9	0.45
3	21	20	1	0.05
4	16	20	16	0.80
5	25	20	25	1.25
6	22	20	4	0.20
7	20	20	0	0.00
8	21	20	1	0.05
9	15	20	25	1.25
Total	200	200		43

$$\chi^2_{cal} = \Sigma \frac{(O_i - e_i)^2}{e_i}$$
$$= 4.3$$

(v) **Conclusion:** Since the calculated value of χ^2, $\chi^2_{cal} = 4.3$ is less than the critical value of $\chi^2_\alpha = 16.22$, we accept null hypothesis. Hence, we conclude that the digits were distributed in equal numbers.

8.8 TEST FOR INDEPENDENCE OF ATTRIBUTES

In the previous topic, we have discussed one way of classification of tables where the observed frequencies are given in a single row. However, some of the events can be analysed to fall into two rows and columns. In such case we have a contingency table.

Suppose we consider two attributes A and B where A is divided into r classes $A_1, A_2, \ldots A_r$ and B is divided into s classes $B_1, B_2, \ldots B_s$.

$r \times s$ contingency table

	B_1	B_2	…	B_j	…	B_s	Total
A_1	O_{11}	O_{12}	…	O_{ij}	…	O_{1s}	(A_1)
A_2	O_{21}	O_{22}	…	O_{2j}	…	O_{2s}	(A_2)
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
A_i	O_{i1}	O_{i2}	…	O_{ij}	…	O_{is}	(A_i)
⋮	⋮	⋮	⋮	⋮		⋮	⋮

A_r	O_{r1}	O_{r2}	…	O_{rj}	…	O_{rs}	(A_r)
Total	(B_1)	(B_2)	…	(B_j)	…	(B_s)	N

In the above table, O_{ij} represents the cell frequency, number of persons possessing both the attributes A_i and B_j where (A_i) is the total number of persons possessing the attribute A_i $(i = 1, 2, \ldots r)$ and (B_j) is the total number of persons possessing the attribute B_j, $j = 1, 2, \ldots s$.

In addition, $\sum_{i=1}^{r}(A_i) = \sum_{j=1}^{s}(B_j) = N$, where N is the total frequency.

Under the null hypothesis that the attributes are independent, the expected frequencies are calculated as follows:

e_{ij} = expected number of persons possessing both the attributes A_i and B_j

$$= \frac{(A_i)(B_j)}{N} \qquad \begin{matrix} i = 1, 2, \ldots r \\ j = 1, 2, \ldots s \end{matrix}$$

Hence, the test statistic is given by

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{s}\frac{(O_{ij} - e_{ij})^2}{e_{ij}}$$

Where O_{ij} is the observed frequency for the contingency table for row i and column j.
e_{ij} is the expected frequency for the contingency table for row i and column j,
follows χ^2 distribution with $\gamma = (r - 1)(s - 1)$ degrees of freedom.

Caution:

- The degrees of freedom are $(n - 1)$ for one way classification of χ^2. One degree of freedom is lost because of the linear constraint $\sum_i O_i = \sum e_i = N$ on the frequencies.
- If any of the population parameters are calculated from the given data and is used for calculating the expected frequencies then we have to subtract one degree of freedom for each parameter calculated.
- If any of the theoretical frequencies are less than 5, then we have to subtract the degrees of freedom lost in pooling these frequencies with the preceding or succeeding frequency.

Worked Out Examples

EXAMPLE 8.22

To test the effectiveness of inoculation against cholera, the following results were obtained:

	Attacked	Not attacked
Inoculated	30	160
Not inoculated	140	460

Does inoculation prevents attack from cholera?

Solution:

	Attacked	Not attacked	Total
Inoculated	30	160	190
Not Inoculated	140	460	600
Total	170	620	790

(i) **Null hypothesis:** Inoculation does not prevent attack from cholera.

(ii) **Alternative hypothesis:** Inoculation prevents attack form cholera.

(iii) **Level of significance:** $\alpha = 5\%$. The critical value of χ^2 for $\gamma = (2 - 1)(2 - 1) = 1$ degree of freedom is $\chi^2_{0.05} = 3.841$.

$$a_{11} = \frac{190 \times 170}{790} = 40.886, \quad a_{12} = \frac{620 \times 190}{790} = 149.114$$

$$a_{21} = \frac{170 \times 600}{790} = 129.114 \quad a_{22} = \frac{620 \times 600}{790} = 470.886$$

Observed frequencies O_i	Expected frequencies e_i	$(O_i - e_i)^2$	$\frac{(O_i - e_i)^2}{e_i}$
30	40.886	118.505	2.898
160	149.114	118.504	0.7947
140	129.114	118.504	0.9178
460	470.886	118.504	0.252
			4.8625

(iv) **Test statistic:** Under H_0, the test statistic, that is,

(v) $$\chi^2 = \Sigma \frac{(O_i - e_i)^2}{e_i}$$
$$= 4.8625$$

(vi) **Conclusion:** Since the calculated value of $\chi^2 = 4.8625$ is greater than the tabulated value of χ^2, $\chi^2_\alpha = 3.841$, we reject null hypothesis and accept alternative hypothesis. Hence, inoculation prevents attack from cholera.

EXAMPLE 8.23

In a certain sample of 2,000 families, 1,400 people are consumers of tea. Out of 1,800 Hindu families, 1,236 families are consumers of tea. State whether there is any significant difference between consumption of tea among Hindu and non-Hindu families.

Solution: The given data is represented in the following table:

	Hindu families	Non-Hindu families	Total
Families consuming tea	1236	164	1400
Families not consuming tea	564	36	600
Total	1800	200	2000

The expected frequencies are calculated as

$$a_{11} = \frac{1400 \times 1800}{2000} = 1260, \quad a_{12} = \frac{200 \times 1400}{2000} = 140$$

$$a_{21} = \frac{1800 \times 600}{2000} = 540, \quad a_{22} = \frac{200 \times 600}{2000} = 60$$

Observed frequencies O_i	Expected frequencies e_i	$(O_i - e_i)^2$	$\frac{(O_i - e_i)^2}{e_i}$
1236	1260	576	0.4571
164	140	576	4.1143
564	540	576	0.0444
36	60	576	9.6
			14.2158

(i) **Null hypothesis:** There is no significant difference between consumption of tea among Hindu and non-Hindu families.

(ii) **Alternative hypothesis:** There is a significant difference between consumption of tea among Hindu and non-Hindu families.

(iii) **Level of significance:** $\alpha = 5\%$. The critical value of χ^2 for $\gamma = (2 - 1)(2 - 1) = 1$ degree of freedom is $\chi^2_{0.05} = 3.841$.

(iv) **Test statistic:** Under H_0, the test statistic is

(v) $$\chi^2 = \frac{\sum (O_i - e_i)^2}{e_i}$$
$$= 14.2158$$

(vi) **Conclusion:** Since the calculated value of $\chi^2 = 14.2158$ is very much greater than critical value, $\chi^2_{0.05} = 3.841$. We reject null hypothesis and accept alternative hypothesis. Hence, there is a significant difference in the consumption of tea among Hindu and non-Hindu families.

EXAMPLE 8.24

A drug is claimed to be effective in curing cold. In an experiment on 160 persons with cold, half of them were given the drug and half of them were given sugar pills. The patients, reactions to the treatment were recorded in the following table:

	Helped	Harmed	No effect
Drug	52	10	20
Sugar pills	44	12	26

Test the hypothesis that the drug is not better than the sugar pills in curing cold.

Solution:

(i) **Null hypothesis:** The drug and sugar pills are equally effective in curing cold.

	Helped	Harmed	No effect	Total
Drug	52	10	20	82
Sugar pills	44	12	26	82
Total	96	22	46	164

The expected frequencies are calculated as follows:

$$a_{11} = \frac{82 \times 96}{164}, \quad a_{12} = \frac{82 \times 22}{164}, \quad a_{13} = \frac{82 \times 46}{164}$$
$$= 48 \qquad = 11 \qquad = 23$$

$$a_{21} = \frac{82 \times 96}{164}, \quad a_{22} = \frac{82 \times 22}{164}, \quad a_{23} = \frac{82 \times 46}{164}$$
$$= 48 \qquad = 11 \qquad = 23$$

Observed frequencies O_i	Expected frequencies e_i	$(O_i - e_i)^2$	$\frac{(O_i - e_i)^2}{e_i}$
52	48	16	0.333
10	11	1	0.0909
20	23	9	0.31034
44	48	16	0.3333
12	11	1	0.0909
26	23	9	0.3913
			2.5263

(ii) **Alternative hypothesis:** The drug and sugar pills are not equally effective in curing cold.

(iii) **Level of significance:** $\alpha = 5\%$. The critical value of χ^2 for $\gamma = (3 - 1)(2 - 1) = 2$ degrees of freedom is $\chi^2_{0.05} = 5.991$.

(iv) **Test statistic:** Under H_0, the test statistic is

$$\chi^2 = \frac{\sum(O_i - e_i)^2}{e_i}$$

$$= 2.5263$$

(v) **Conclusion:** Since the calculated value of $\chi^2 = 2.5263$ is less than the critical value of $\chi^2 = 5.991$, we accept null hypothesis.

Hence, the drug and sugar pills are equally effective in curing cold.

EXAMPLE 8.25

1,000 families were selected at random, in a city to test the belief that the high income group families usually send their children to public schools and the low income families often send their children to government schools. The following results were obtained:

	School		
Income	Public	Government	Total
Low	370	430	800
High	130	70	200
Total	500	500	1000

Test whether income and type of schooling are independent.

Solution:

(i) **Null hypothesis:** The income and type of schooling are independent.
The expected frequencies are calculated as follows:

$$a_{11} = \frac{800 \times 500}{1000} = 400, \qquad a_{12} = \frac{500 \times 800}{1000} = 400$$

Observed frequencies O_i	Expected frequencies e_i	$(O_i - e_i)^2$	$\frac{(O_i - e_i)^2}{e_i}$
370	400	900	2.25
430	400	900	2.25
130	100	900	9
70	100	900	9
			22.5

(ii) **Alternative hypothesis:** The income and type of schooling are not independent.

(iii) **Level of significance:** $\alpha = 5\%$. The critical of χ^2 for $\gamma = (r-1)(s-1) = 1$ degree of freedom $\chi^2_{0.05} = 3.841$.

(iv) **Test statistic:** Under H_0, the test statistic is

$$\chi^2 = \Sigma\Sigma \frac{(O_{ij} - e_{ij})^2}{e_{ij}}$$

$$= 22.5$$

(v) **Conclusion:** Since the calculated value of $\chi^2 = 22.5$ is very much greater than the critical value of χ^2, $\chi^2 = 3.841$, we reject null hypothesis and accept alternative hypothesis. Hence, the income and type of schooling are not independent.

EXAMPLE 8.26

In the accounting department of a bank, 100 accounts are selected at random and examined for errors. Suppose the following results were obtained:

No. of errors	0	1	2	3	4	5	6
No. of accounts	35	40	19	2	0	2	2

Can it be concluded that the errors are distributed according to the Poisson probability law?

Solution: We have to find expected frequencies by fitting Poisson distribution.

No. of errors, x_i	0	1	2	3	4	5	6
No. of accounts, f_i	35	40	19	2	0	2	2
$f_i x_i$	0	40	38	6	0	10	12

$$\Sigma f_i x_i = 106$$

$$N = \Sigma f_i = 100$$

$$\text{Mean, } \bar{x} = \frac{\Sigma f_i x_i}{N} = \frac{106}{100} = 1.06$$

Mean of Poisson distribution = Mean of the given distribution

$$\lambda = 1.06$$

The probabilities are calculated as follows:

$$P(x+1) = \left(\frac{\lambda}{x+1}\right) P(x), \qquad x = 0,1,2,3,4,5,6$$

Their expected frequencies are calculated as $e_i = f(x_i) = NP(x_i)$

$$P(0) = e^{-\lambda} = e^{-1.06} = 0.3464$$

$$f(0) = N \cdot P(0) = 100(0.3464) = 34.64 \cong 35$$

$$P(1) = (1.06)\, P(0) = 0.3671$$

$$f(1) = N \cdot P(1) \;=\; 36.718 \cong 37$$

$$P(2) = \left(\frac{1.06}{2}\right) P(1)$$

$$= 0.1945$$

$$f(2) = 19.45 \cong 19$$

$$P(3) = \left(\frac{1.06}{3}\right) P(2)$$

$$= 0.06872$$

$$f(3) = 100P(3) = 6.872 \cong 7$$

$$P(4) = \left(\frac{1.06}{4}\right) P(3)$$

$$= 0.01821$$

$$F(4) = 100\, P(4)$$

$$= 1.82 = 1.82 = 2$$

$$P(5) = \left(\frac{1.06}{5}\right) P(4) = 0.00386$$

$$f(5) = 100P(5) = 0.386 \sim 0$$

$$P(6) = \left(\frac{1.06}{6}\right) P(5)$$

$$= 0.000409$$

$$f(6) = 0.409 \sim 0$$

Observed frequency O_i	Expected frequency e_i	$(O_i - e_i)^2$	$\frac{(O_i - e_i)^2}{e_i}$
35	35	0	0
40	37	9	0.2432
19	19	0	0
2	7		
0	2		1
2 (6)	0 (9)	(9)	1.2432
2	0		
	0		

(i) **Null hypothesis:** The errors are according to Poisson law.

(ii) **Alternative hypothesis:** They are not according to Poisson law.

(iii) **Level of significance:** $\alpha = 5\%$, χ^2_α for $\gamma = 0 - 1 = 7 - 1 - 1 - 3 = 0.455$

(one degree of freedom is lost for evaluation of $\bar{x}$, 1 for linear constraint $\sum O_i = \sum e_i = 3$ degrees of freedom are lost because of pooling the last four e_i which are less than 5).

(iv) **Test statistic:** Under H_0, $x^2 = \sum \frac{(O_i - e_i)^2}{e_i}$

$= 5.1002$

(v) **Conclusion:** Since $\chi^2_{cal} > x^2_\alpha$, we reject H_0 and accept H_1.

Hence the errors are according to Poisson law.

Work Book Exercises

1. A random sample of size 16 has 53 as mean. The sum of squares of the deviations taken from mean is 135. Can this sample be regarded as taken from the population having 56 as mean?
2. The average breaking strength of steel rods is specified to be 18.5 thousand lbs. To test this, sample of 14 rods were tested. The mean and standard deviation obtained were 17.85 and 1.955, respectively. Is the result of the experiment significant? [**Ans.:** $t = -1.20$]
3. A fertilizer mining machine is set to give 12 kg of nitrate for every quintal bag of fertilizer. Ten 100 kg bags are examined. The percentages of nitrate are 11, 14, 13, 12, 13, 12, 13, 14, 11, and 12. Is there a reason to believe that the machine is defective?
4. The yield of a crop from six test plots is 2.75, 5.25, 4.50, 2.50, 4.25, and 3.25 tonnes per hectare. Test at 5% level of significance whether this supports the contention that the true average yield for this kind of crop is 3.50 tonnes per unit.
5. The life time of electric bulbs for a random sample of 10 from a large consignment gave the following data:

Item	1	2	3	4	5	6	7	8	9	10
Life in 1,000 hours	4.2	4.6	3.9	4.1	5.2	3.8	3.9	4.3	4.4	5.6

Can we accept the hypothesis that average life time of bulbs is 4,000 hours? [**Ans.:** $t = 2.148$]

6. The average length of time for students to register for summer classes at a certain college has been 50 minutes with a standard deviation of 10 minutes. A new registration procedure using modern computing machines is being tried. If a random sample of 12 students had an average registration time of 42 minutes with standard deviation of 11.9 minutes under the new system, test the hypothesis that the population mean has not changed, using 0.05 level of significance.
7. A random sample of 8 envelopes is taken from a letter box of a post office and their weights in grams are found to be 12.1, 11.9, 12.4, 12.3, 11.9, 12.1, 12.4, and 12.1. Does this sample indicate at 1% level that the average weight of envelopes received at that post office is 12.35 grams?
8. The following data about the life of two brands of bulbs is as follows:

	Mean life	Standard deviation	Size of sample
Brand A	2000 hours	250 hours	12
Brand B	2230 hours	300 hours	15

Is there a significant difference in the two bulbs?

9. The incomes of a random sample of engineers in industry *A* are ₹630, 650, 680, 690, 710, and 720 per month. The incomes of a similar sample from industry *B* are ₹610, 620, 650, 660, 690, 700, 710, 720, and 730 per month. Discuss the validity of the suggestion that industry *A* pays its engineers much better than industry *B*. [**Ans.:** $t = 0.099$]

10. The sales data of an item in six shops before and after a special promotional campaign are as follows:

Shops	A	B	C	D	E	F
Before campaign	53	28	31	48	50	42
After campaign	58	29	30	55	56	45

Can the campaign be judged to a success at 5% level of significance?

11. A reading test is given to an elementary school class that consists of 12 Anglo-American children and 10 Mexican-American children. The results of the test are as follows:

Anglo-American	Mexican-American
$\bar{x}_1 = 74$	$\bar{x}_2 = 70$
$S_1 = 8$	$S_2 = 10$

Is the difference between the means of the two groups significant at 5% level of significance?

12. Two laboratories *A* and *B* carry out independent estimates of fat content in ice-cream made by a firm. A sample is taken from each batch. The fat content is recorded and is as follows:

Batch No.	1	2	3	4	5	6	7	8	9	10
Lab *A*	7	8	7	3	8	6	9	4	7	8
Lab *B*	9	8	8	4	7	7	9	6	6	6

Is there a significant difference between the mean fat content obtained by the two laboratories *A* and *B*?

13. Samples of sales in similar shops in towns *A* and *B* regarding a new product yielded the following information:

Town A	$\bar{x}_i = 3.45$	$\Sigma x_i = 38$	$\Sigma x_i^2 = 228$	$n_1 = 11$
Town B	$\bar{y}_j = 4.44$	$\Sigma y_j = 40$	$\Sigma y_j^2 = 222$	$n_2 = 9$

Is there any evidence of difference in sales in the two towns? $\Sigma x_i = \Sigma(x_i - \bar{x}_i)$, $\Sigma x_i^2 = \Sigma(x_i - \bar{x}_i)^2$, $\Sigma y_j = \Sigma(y_j - \bar{y}_j)$, and $\Sigma y_j^2 = \Sigma(y_j - \bar{y}_j)^2$.

14. The average number of articles produced by two machines per day is 200 and 250 with standard deviations 20 and 25, respectively on the basis of records of 25 days production. Can you regard both the machines equally efficient at 1% level of significance?

15. An IQ test was administered to 5 persons before and after they were trained. The results are as follows:

Candidates	1	2	3	4	5
IQ before training	110	120	123	132	125
IQ after training	120	118	125	136	121

Test whether there is any change in the IQ after the training programme.

16. The increase in weight in kilograms in particular period of 10 students of a certain age group of a high school fed with nourishing food "COMPLAN" were observed as 5, 2, 6, −1, 0, 4, 3, −2, 1, and 4. 12 students of the same age group, but of another high school were fed with another nourishing food "ASTRA" and the increase in weight in kilograms in the same period were observed as 2, 8, −1, 5, 3, 0, 6, 1, −2, 0, 4, and 5. Test whether the two foods COMPLAN and ASTRA differ significantly as regards the effect on the increase in weight.

17. Two laboratories carry out independent estimates of a particular chemical in a medicine produced by a certain firm. A sample is taken from each batch.

No. of samples	10
Mean value of the difference of estimates	0.6
Sum of squares of the differences from their means	20

Is the difference significant?

18. A drug was administered to 10 patients, and the increments in their blood pressure were recorded to be 6, 3, − 2, 4, − 3, 4, 6, 0, 3, and 2. Is it reasonable to believe that the drug has no effect on change of blood pressure?

19. Two random samples were drawn from the normal populations are as follows:

Sample I	20	16	26	27	23	22	18	24	25	19	30
Sample II	27	33	42	35	32	34	38	28	41	43	37

Test whether the two populations have same variance.

20. The following data gives the prices in rupees of a certain commodity in a sample of 15 shops selected at random from a city *A* and those in a sample of 13 shops from another city *B*.

City *A*	7.41	7.77	7.44	7.40	7.38	7.93	7.58	8.28	7.23	7.52	7.82	7.71	7.84	7.63	7.68
City *B*	7.08	7.49	7.42	7.04	6.92	7.22	7.68	7.24	7.74	7.81	7.28	7.43	7.47	–	–

Is it reasonable to say that the variability of prices in the two cities is the same?

21. The following data relates to a random sample of government employees in two Indian states of the Indian Union:

	State I	State II
Sample size	16	25
Mean monthly income	440	460
Sample variance	40	42

Test the hypothesis that the variances of the two populations are equal.

22. The time taken by workers in performing a job by method I and method II are as follows:

Method I	20	16	26	27	23	22	–
Method II	27	33	42	35	32	34	38

Do these data show that the variances of time distribution in a population from which these samples are drawn do not differ significantly? [**Ans.:** $F = 1.37$]

23. In a laboratory experiment two random samples gave the following result:

Sample	Size	Sample mean	Sum of squares of deviations from the mean
1	10	15	90
2	12	14	108

Test the equality of sample variance at 5% level of significance.

24. The Nicotine content in milligrams of two samples of tobacco was found to be as follows:

Sample A	24	27	26	21	25	–
Sample B	27	30	28	31	22	36

Can it be regarded that the two samples come from the same normal population?

25. A typist in a company commits the following number of mistakes per page in typing 432 pages:

No. of mistakes per page	0	1	2	3	4	5	Total
No. of pages	223	142	48	15	4	0	432

Does this information verify that the mistakes are distributed according to the Poisson law?

26. The number of road accidents per week in a certain area is as follows: 12, 8, 20, 2, 17, 10, 15, 6, 9, 4. Are these frequencies in agreement with the belief that accident conditions were the same during the 10-week period?

27. Four dice were thrown 112 times and the number of times 1, 3, or 5 was thrown were as follows:

No. of times dice 1, 3 or 5 was thrown	0	1	2	3	4
Frequency	10	25	40	30	7

Are the dice all fair?

28. Fit a Poisson distribution to the following data and test the goodness of fit:

X	0	1	2	3	4	5	6
F	275	72	30	7	5	2	1

29. Two researchers adopted different sampling techniques while investigating the same group of students to find the number of students falling in different intelligence levels. The results are as follows:

Researcher	No. of students in each level				
	Below average	Average	Above average	Medium	Total
X	86	60	44	10	200
Y	40	33	25	2	100
Total	126	93	69	12	300

Would you say that the sampling techniques adopted by the two researchers are significantly different?

30. The following data shows whether a certain vaccination prevents a certain disease or not, an experiment was conducted and the following figures in various chances were as follows:

	Vaccinated	Not vaccinated	Total
Attacked	69	10	79
Not attacked	91	30	121
Total	160	40	200

Use χ^2-test to find if there is dependence in the two attributes.

31. For the data in the following table, test for independence of person's ability in Mathematics and interest in Economics:

		Ability in Mathematics		
		Low	Average	High
Interest in Economics	Low	63	42	15
	Average	58	61	31
	High	14	47	29

32. A sample of 320 units of a manufactured product classified according to the quality of the product and the production shift is as follows:

	Quality	
	Good	Bad
Day shift	150	20
Night shift	105	45

Find whether quality depends on shift.

33. In a survey of 200 boys, of which 75 were intelligent, 40 had skilled fathers while 85 of the unintelligent boys had unskilled fathers. Do these figures support the hypothesis that skilled fathers have intelligent boys?

34. A drug is claimed to be effective in curing cold. In an experiment on 164 persons with cold, half of them were given drug, and half were given sugar pills. The treatments are recorded as follows:

	Helped	Harmed	No effect
Drug	52	10	20
Sugar pills	44	12	26

Test the hypothesis that the drug is not better than the sugar pills in curing colds.

35. A company has two factories one located in Delhi and another in Mumbai. It is interested to know whether its workers are satisfied with their jobs or not at both the places. To get proper and reliable information it has undertaken a survey at both the factories and the data obtained are as follows:

	No. of workers by degree of satisfaction		
	Delhi	Mumbai	Total
Fully satisfied	50	70	120
Moderately Satisfied	90	110	200
Moderately dissatisfied	160	130	290
Fully dissatisfied	200	190	390
Total	500	500	1000

DEFINITIONS AT A GLANCE

***t*-statistic:** A statistic which is used to test for single mean, difference of means, and paired differences for small samples.

***F*-distribution:** A continuous distribution that has two parameters. It is used mainly to test hypothesis concerning variances.

χ^2-distribution: A distribution with degrees of freedom as the only parameter. It is skewed to the right for small degrees of freedom, looks like a normal curve for large degrees of freedom.

Contingency Table: A table having rows and columns where in each row corresponds to level of one variable and each column to another.

Degrees of Freedom: The number of elements that can be chosen freely.

FORMULAE AT A GLANCE

- Test for single mean

$$t = \frac{\bar{x} - \mu}{\frac{S}{\sqrt{n}}}$$

where $\bar{x} = \frac{\sum x_i}{n}$, $S = \frac{1}{n-1}\sum (x_i - \bar{x})^2$ with $\gamma = n - 1$ degrees of freedom

- *t*-test for difference of means

$$t = \frac{\bar{x} - \bar{y}}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \text{ where } \bar{x} = \frac{\sum x_i}{n_1},\ \bar{y} = \frac{\sum y_i}{n_2}$$

$$s = \frac{\sum (x_i - \bar{x})^2 + \sum (y_j - \bar{y})^2}{n_1 + n_2 - 2} = \frac{n_1 s_1^{\,2} + n_2 s_2^2}{n_1 + n_2 - 2}$$

Follows *t*-distribution with $\gamma_1 = n_1 + n_2 - 2$ degrees of freedom.

- Paired *t*-test for difference of means

$$t = \frac{\bar{d}}{\frac{S}{\sqrt{n}}} \text{ where } \bar{d} = \frac{\sum di}{n},\ d_i = x_i - y_i$$

$$S = \frac{1}{n-1}\sum (d_i - \bar{d})^2$$

Follows student's *t*-distribution with (n − 1) degrees of freedom.

- *F*-test for equality of population variances
- $F = \frac{S_x^{\,2}}{S_y^{\,2}}$ where $S_x^{\,2} = \frac{1}{n_1 - 1}\sum (x_i - \bar{x})^2$

$$S_y^2 = \frac{1}{n_2 - 1}\Sigma(y_j - \bar{y})^2$$

Follows F-distribution with $(g_1, g_2) = (n_1 - 1, n_2 - 1)$ degrees of freedom.

- χ^2-test for goodness of fit,

  $\chi^2 = \sum_{i=1}^{n} \frac{(O_i - e_i)^2}{e_i}$, where O_i is the observed frequencies, e_i is expected frequencies. Follows χ^2-distribution with $\gamma = n - 1$ degrees of freedom.

- χ^2-test for $r \times s$ contingency table is

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{(O_{ij} - e_{ij})^2}{e_{ij}}$$

OBJECTIVE TYPE QUESTIONS

1. The statistic used to test if there is a significant difference between the sample mean $\bar{x}$ and population mean μ for small n,

   (a) z-test (b) t-test
   (c) F-test (d) none of these

2. t-test for difference of means is given by $t =$

   (a) $\frac{(\bar{x} - \bar{y}) - (\mu_x - \mu_y)}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$, where $S = \frac{n_1 S_1^2 + n_2 S^2}{n_1 + n_2 - 2}$ (b) $\frac{\bar{x} - \bar{y}}{S\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$, where $S = \frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2 - 2}$

   (c) $\frac{S\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}{\bar{x} - \bar{y}}$, where $S = \frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2 - 2}$ (d) none

3. The degrees of freedom used for testing the difference of means for $n_1 = 12$, $n_2 = 10$ is $\gamma =$

   (a) 22 (b) 20
   (c) 19 (d) none

4. The test statistic used for testing difference of means when sample observations are not independent but paired together then, $t =$

   (a) $\frac{d_i}{\frac{s}{\sqrt{n}}}$, where $di = x_i - y_i$ (b) $\frac{\bar{d}}{\frac{s}{\sqrt{n}}}$, where $\bar{d} = \frac{\Sigma d_i}{n}$, $di = x_i - y_i$

   (c) $\frac{\bar{d}}{\frac{s}{\sqrt{n}}}$, where $\bar{d} = \frac{\Sigma d_i}{n}$, $d_i = x_i - y_i$ (d) none

5. The statistic used to test $H_0: \sigma_x^2 = \sigma_y^2$ against $H_1: \sigma_x^2 \neq \sigma_y^2$ is
   (a) F-test (b) t-test
   (c) χ-test (d) none

6. If $F_{0.05}(4, 5) = 5.19$, then $F_{0.95}(5, 4) =$
   (a) 0.5 (b) 0.19
   (c) 1.95 (d) none

7. The χ^2-statistic used to test independence of attributes is given by $\chi^2 =$
   (a) $\Sigma \dfrac{(O_i - e_i)}{e_i}$ (b) $\Sigma \dfrac{O_i - e_i}{O_i}$
   (c) $\Sigma \dfrac{(O_i - e_i)^2}{e_i}$ (d) none

8. In 4×5 contingency table, the degrees of freedom obtained to find χ^2 is
   (a) $(4-1) \times (5-1)$ (b) $4 + 5 - 2$
   (c) $(4-1) + (5-1)$ (d) none

9. The expected frequencies for $r \times s$ contingency table are calculated using one of the following formulas given $\sum_{i=1}^{r} (A_i) = \sum_{j=1}^{s} B_j = N$ is $e_{ij} =$
   (a) $(A_i)\,(B_i)$ (b) $\dfrac{(A_i)(B_i)}{N}$
   (c) $N(A_i)\,(B_i)$ (d) none

10. $t_{1-\alpha} =$ _____
    (a) t_α (b) $-t_\alpha$
    (c) $t_{\alpha-1}$ (d) none

11. $F_{1-\alpha}(\gamma_1, \gamma_2) =$ _____
    (a) $F_\alpha(\gamma_2, \gamma_1)$ (b) $F_\alpha(\gamma_1, \gamma_2)$
    (c) $\dfrac{1}{F_\alpha(\gamma_1, \gamma_2)}$ (d) none

ANSWERS

1. (b) 2. (a) 3. (b) 4. (c) 5. (a) 6. (b) 7. (c) 8. (a)
9. (b) 10. (b) 11. (a)

9 Estimation

Prerequisites

Before you start reading this unit, you should:

- Know normal tables and calculating *Z* from the tables
- Know the previous units on sampling theory
- Know calculation of mean and standard deviation for a data

Learning Objectives

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

- Know the necessity of estimation and differentiate between point and interval estimates
- Know the criteria of a good estimator
- Know the procedures involved in constructing confidence intervals for sample means, proportions, difference of means, and difference of proportions
- Determine the sample size in estimation
- Determine the maximum error in estimating sample mean or sample proportion

INTRODUCTION

The theory of statistical inference consists of those methods by which one makes inferences or generalizations about a population. In Units 7 and 8 we did not attempt to estimate the population parameters, but we tried to arrive at a correct decision about a pre-stated hypothesis.

A parameter may have more than one estimator. An estimator is not expected to estimate the population parameter without error. The estimation can be done in the following two ways:

(i) Point estimation
(ii) Interval estimation

9.1 POINT ESTIMATION

A point estimate of some population parameter θ is a single value $\hat{\theta}$ of a statistic $\hat{\theta}$.

For example, the value $\bar{X}$ which is computed from a sample of size n is a point estimate of the population parameter μ, that is, $\bar{X}$ is the point estimate for the population mean μ. Each sample may give an estimate for the population parameter. However, of all the estimates, which one would be the better one.

Hence there should be some desirable properties of a good estimator which make us to choose one estimator rather than another.

Caution:

- The statistic whose distribution concentrates as closely as possible near the true value of the parameter is called best estimate.

- The best estimate would be one that falls nearest to the true value of the parameter to be estimated.
- The estimating functions are called as estimators.

9.2 CHARACTERISTICS OF ESTIMATORS

A good estimator should possess the following characteristics:

Consistency

Let n be the size of the random sample. An estimator T_n based on this random sample is said to be consistent estimate of $\gamma(\theta)$ if T_n converges to $\gamma(\theta)$ in probability, that is, if for every $\in > 0, \eta > 0$, these exists a positive integer $n \geq m(\in, \eta)$ such that $p\left[|T_n - \gamma(\theta)| < \in\right] \to 1$ as $n \to \infty$.

Caution:

Consistency is concerning the behaviour of an estimate for large values of the sample size n, that is, as $n \to \infty$.

We have seen that the above property can be checked for infinitely large samples. Now we move to next property which can be checked for finite 'n'.

Unbiasedness

An estimator T_n is said to be unbiased estimator of $\gamma(\theta)$ if mean of the sampling distribution of T_n is equal to the parameter estimated.

$$E(T_n) = \gamma(\theta) \text{ for all } \theta \in \theta$$

For example, now let us show that the sample mean and sample variance (S^2) are unbiased estimators of population mean and population variances, respectively.

Let $x_1, x_2, \ldots x_n$ be a random sample of size n from a large population $X_1, X_2, \ldots X_N$ of size N with mean μ and variance σ^2.

The sample mean and variance are given by

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

$$E(\bar{x}) = E\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right]$$

$$= \frac{1}{n}\sum_{i=1}^{n} E(x_i) \tag{9.1}$$

Each x_i is a sample member from the population $X_1, X_2, \ldots X_N$. Hence, it can take any one of the values $X_1, X_2, \ldots X_N$ each with equal probability of $\frac{1}{N}$.

$$\therefore E(x_i) = \frac{1}{N}X_1 + \frac{1}{N}X_2 + \cdots + \frac{1}{N}X_N$$

$$= \frac{1}{N}(X_1 + X_2 + \cdots + X_N)$$

$$= \mu$$

Substituting this value in (9.1) we get,

$$E(\bar{x}) = \frac{1}{n}\sum_{i=1}^{n}\mu$$

$$= \frac{1}{n}(n\mu) = \mu$$

$$\therefore E(\bar{x}) = \mu$$

Hence, we can say that the sample mean $\bar{x}$ is an unbiased estimate of the population mean μ.

Similarly we shall prove that sample variance (S^2) is also unbiased estimate of population variance.

$$\text{Now } E(s^2) = E\left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right]$$

$$= \frac{1}{n}E\left[\sum_{i=1}^{n}(x_i - \mu) - (\bar{x} - \mu)\right]^2$$

$$= \frac{1}{n}E\left[\sum(x_i - \mu)^2 - 2(\bar{x} - \mu)\sum_{i=1}^{n}(x_i - \mu) + n(\bar{x} - \mu)^2\right]$$

$$= \frac{1}{n}E\left[\sum(x_i - \mu)^2 - n(\bar{x} - \mu)^2\right]$$

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

However, we know that $\frac{1}{n}E(x_i - \mu)^2 = \sigma x_i^2 = \sigma^2$ and $E(\bar{x} - \mu)^2 = \sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}$

$$\therefore E(s^2) = \sigma^2 - \frac{\sigma^2}{n} = \left(1 - \frac{1}{n}\right)\sigma^2$$

Hence, $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ is an estimate of population variance.

$$ns^2 = (n-1)S^2 \Rightarrow s^2 = \left(1 - \frac{1}{n}\right)S^2$$

For large samples when $n \to \infty$, we have $s^2 \to S^2$.

Now let us move on to the next property of estimators. We can find two unbiased estimators which are consistent, but now do we choose a better among them. The rent property that is being dealt answers this question.

Efficient Estimators

A criterion which is based on the variances of the sampling distributions of the estimators is known as efficiency.

Let T_1 and T_2 be two consistent estimators of a parameter θ such that variance of T_1 is less than that of T_2, that is, $V(T_1) < V(T_2)$ for all n. Then T_1 is said to be more efficient estimator than T_2 for all sample sizes.

If in a class of consistent estimators for a parameter there exists one whose sampling variance is less than that of all other estimators, then such an estimator is said to be most efficient estimator.

Efficiency

Let T, T_1, $T_2 \ldots$, T_n be the estimators of $\gamma(\theta)$ and variance of T is the minimum compared to variances of all other estimators then

Efficiency of T_i, $E_i = \frac{\text{var}(T)}{\text{var}(T_i)}, i = 1, 2, \ldots n$

E_i is always less than 1.

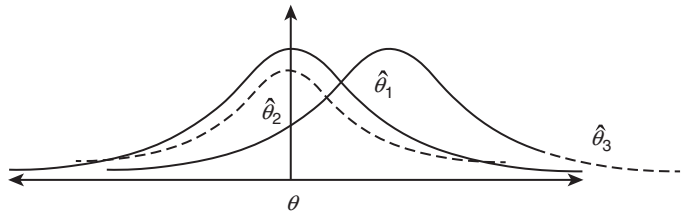
$$\text{(i.e.,) } E_i \leq 1,\ i = 1, 2, \ldots n.$$

Another definition that can be drawn from the above is as follows:

Minimum Variance Unbiased (MVU) Estimators
Let T be a statistic based on a random sample of size n such that

(i) T is an unbiased estimator of $\gamma(\theta)$, for all $\theta \in \boldsymbol{\theta}$.
(ii) Var (T) is the smallest among all unbiased estimators of $\gamma(\theta)$, then T is called the minimum variance unbiased estimator of $\gamma(\theta)$.

The following figure gives three different estimators which are unbiased, since their discriminations are centered along θ.



The sampling distribution of three estimators $\hat{\theta}_1$, $\hat{\theta}_2$, and $\hat{\theta}_3$ are given.
Here $\hat{\theta}_1$ is more efficient estimator.

Sufficiency

An estimator is said to be sufficient for a parameter, if it contains all the information in the sample regarding the parameter, that is, an estimator T based on a random sample size n is said to be sufficient if the sample $x_1, x_2, \ldots x_n$ is drawn from a population with density $f(x, \theta)$ such that the conditional distribution of $x_1, x_2, \ldots x_n$ given T is independent of θ.

9.3 INTERVAL ESTIMATION

So far we have been discussing point estimation, the criteria for a good estimator. Now we shall introduce the concept of interval estimation.

Even the most efficient estimator may not estimate the population parameter exactly. As the accuracy increases with large samples, but the point estimate may not exactly be equal to the population parameter. Hence, it is desirable to determine an interval within which we would expect to find the value of the parameter and such an interval is called interval of estimation. An interval estimate of a population parameter θ is an interval of the form $\hat{\theta}_L < \theta < \hat{\theta}_U$ where $\hat{\theta}_L$ and $\hat{\theta}_U$ depend on the following:

(i) The value of the statistic $\hat{\theta}$
(ii) On the sampling distribution of $\hat{\theta}$

9.4 CONFIDENCE INTERVAL

Now, we shall develop the idea of confidence interval and confidence limits.

If we can find $\hat{\theta}_L$ and $\hat{\theta}_U$ such that

$$P(\hat{\theta}_L < \theta < \hat{\theta}_u) = 1 - \alpha$$

For $0 < \alpha < 1$, then we have a probability of selecting a random sample that will produce an interval containing the parameter θ. Such an interval $(\hat{\theta}_L, \hat{\theta}_U)$ which is computed from the selected sample is called a $(1 - \alpha)100\%$ confidence interval, the fraction $(1 - \alpha)$ is called confidence coefficient. The end points of the interval $\hat{\theta}_L$ and $\hat{\theta}_U$ are called confidence limits.

For example, if $\alpha = 0.05$, then we get $(1 - \alpha)100\% = 95\%$ confidence interval for any parameter.

It is to be noted here that though point estimation and interval estimation represent different approaches, but they are related as confidence interval estimators and are based on point estimators in gaining information regarding the parameter.

Confidence Interval for Single Mean μ

From the above theory let us develop the theory of interval for mean μ.

Let the sample be drawn from normal population or let the size of the sample be sufficiently large.

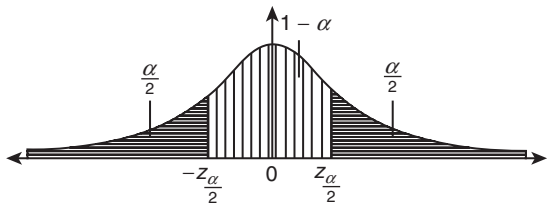
According to central limit theorem, the sampling distribution of $\bar{x}$ is expected to be approximately normally distributed with $\mu_{\bar{x}} = \mu$ and $\sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}}$

Then we know that

$$P(-z_{\alpha/2} < z < z_{\alpha/2}) = 1 - \alpha \tag{9.2}$$

From the theorem, $z = \dfrac{\bar{x} - \mu}{\dfrac{\sigma}{\sqrt{n}}}$

Substituting this in (9.2) we get



$$P\left(-z_{\frac{\alpha}{2}} < \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} < z_{\frac{\alpha}{2}}\right) = 1 - \alpha$$

$$P\left(-z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} < \bar{x} - \mu < \frac{\sigma}{\sqrt{n}} z_{\frac{\alpha}{2}}\right) = 1 - \alpha$$

$$P\left(-z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} < \mu < z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

(i.e.,) $$P\left(\bar{x} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} < \mu \ \bar{x} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

{By multiplying by –1 and subtracting $\bar{x}$ from the inequality}

Hence, a $(1 - \alpha)100\%$ confidence interval for μ of a random sample of size n with mean $\bar{x}$ which is drawn from a population with known variance σ^2 is given by

$$\left(\bar{x} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right)$$

where $z_{\frac{\alpha}{2}}$ is the z-value leaving an area of $\dfrac{\alpha}{2}$ to the right.

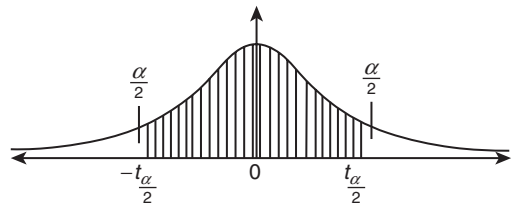
Confidence Interval for μ When Variance is Unknown

In the previous case the variance is known. Suppose that the variance is unknown, that is, if we have a random sample from a normal distribution, then the random variable $t = \dfrac{\bar{x} - \mu}{\dfrac{s}{\sqrt{n}}}$ has a student's

t-distribution with $(n-1)$ degrees of freedom. Here, S is the sample standard deviation. This t can be used to construct confidence interval for μ. The procedure is the same as that derived in case I, but σ is replaced by sample standard deviation S and normal distribution is replaced by t-distribution.

$$\therefore P(-t_{\frac{\alpha}{2}} < t < t_{\frac{\alpha}{2}}) = 1-\alpha$$

$$P\left(-t_{\frac{\alpha}{2}} < \frac{\bar{x}-\mu}{\frac{s}{\sqrt{n}}} < t_{\frac{\alpha}{2}}\right) = 1-s\alpha$$



(i.e.,) $P\left(\bar{x}-t_{\frac{\alpha}{2}}\frac{S}{\sqrt{n}} < \mu < \bar{x}+t_{\frac{\alpha}{2}}\frac{S}{\sqrt{n}}\right) = 1-\alpha$

where $t_{\frac{\alpha}{2}}$ is the t-value with $(n-1)$ degrees of freedom above which an area of $\frac{\alpha}{2}$ is observed.

Hence, the $(1-\alpha)100\%$ confidence interval for μ given the random sample is of size n with mean $\bar{x}$ and standard deviation S is given by

$$\bar{x}-t_{\frac{\alpha}{2}}\frac{S}{\sqrt{n}} < \mu < \bar{x}+t_{\frac{\alpha}{2}}\frac{S}{\sqrt{n}}$$

Confidence Interval for Estimating Proportion

Consider a binomial experiment with n trials. The presence of an attribute in a sampled unit is called success and its absence is called failure. Let X represents the number of successes in n trials. The proportion $\hat{P}$ is given by $\hat{P} = \frac{x}{n}$. By central limit theorem for very large n, $\hat{P}$ is approximately distributed with mean

$$\mu_{\hat{P}} = E(\hat{P}) = E\left(\frac{x}{n}\right) = \frac{1}{n}E(x)$$

Since X is the number of success in n independent trials $E(X) = nP$ and $V(x) = nPQ$, we have

$$\mu_{\hat{P}} = \frac{1}{n}nP = P$$

where P is constant probability of success for each trial and $Q = 1 - P$ and variance of $\hat{P}$ is

$$\sigma_{\hat{p}}^2 = \sigma_{\frac{X}{n}}^2 = V\left(\frac{X}{n}\right) = \frac{1}{n^2}V(X)$$

$$= \frac{1}{n^2}(nPQ) = \frac{PQ}{n}$$

Hence, we can assert with $100(1-\alpha)$ % confidence that $P(-z_{\frac{\alpha}{2}} < z < z_{\frac{\alpha}{2}}) = 1-\alpha$

where $z = \frac{\hat{P}-E(\hat{P})}{\sqrt{V(\hat{P})}}$

and $z_{\frac{\alpha}{2}}$ is the value of the standard normal variate above which we can find an area $\frac{\alpha}{2}$. Hence we get

$$P\left(-z_{\frac{\alpha}{2}} < \frac{\hat{P}-P}{\sqrt{\frac{PQ}{n}}} < z_{\frac{\alpha}{2}}\right) = 1-\alpha$$

$$P\left(-z_{\frac{\alpha}{2}}\sqrt{\frac{PQ}{n}} < \hat{P}-P < z_{\frac{\alpha}{2}}\sqrt{\frac{PQ}{n}}\right) = 1-\alpha$$

$$P\left(\hat{P}-z_{\frac{\alpha}{2}}\sqrt{\frac{PQ}{n}} < P < \hat{P}+z_{\frac{\alpha}{2}}\sqrt{\frac{PQ}{n}}\right) = 1-\alpha$$

If P is unknown, then the sample proportion $\hat{P} = \frac{x}{n}$ can be used, as the point estimate of the parameter P.

Hence $100(1-\alpha)\%$ confidence interval for P is

$$\hat{P}-z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{P}\hat{Q}}{n}} < P < \hat{P}+z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{P}\hat{Q}}{n}}$$

where $\hat{P}$ is the proportion of success in a random sample of size n and $\hat{Q} = 1-\hat{P}$ and $z_{\frac{\alpha}{2}}$ is the value of z leaving an area of $\frac{\alpha}{2}$ to the right.

Confidence Interval for Difference Between Two Proportions

This is used when we are interested to estimate the difference between two binomial parameters p_1 and p_2. Let there be two independent samples of sizes n_1 and n_2 drawn from two binomial populations whose means are n_1p_1 and n_2p_2 and variances are $n_1p_1q_1$ and $n_2p_2q_2$.

Let $\frac{x_1}{n_1}$ and $\frac{x_2}{n_2}$ be the proportions of successes denoted by $\hat{p}_1 = \frac{x_1}{n_1}$ and $\hat{p}_2 = \frac{x_2}{n_2}$.

The confidence interval for $p_1 - p_2$ can be established by considering the sampling distribution of $\hat{p}_1 - \hat{p}_2$ each of which are normally distributed with means p_1 and p_2 and variances $\frac{p_1q_1}{n_1}$ and $\frac{p_2q_2}{n_2}$.

$\hat{p}_1 - \hat{p}_2$ will be approximately normally distributed with mean $\mu_{\hat{p}_1-\hat{p}_2} = p_1 - p_2$ and variance $\sigma^2_{\hat{p}_1-\hat{p}_2} = \frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}$.

Hence, we can assert with $100(1-\alpha)\%$ confidence that $P(-z_{\frac{\alpha}{2}} < z < z_{\frac{\alpha}{2}}) = 1-\alpha$ where

$$z = \frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{\frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}}}$$

$$\therefore P\left[-z_{\frac{\alpha}{2}} < \frac{\hat{p} - \hat{p}_2 - (p_1 - p_2)}{\sqrt{\frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}}} < z_{\frac{\alpha}{2}}\right] = 1-\alpha$$

$$P\left[-z_{\frac{\alpha}{2}}\sqrt{\frac{p_1q_1}{n_1}+\frac{p_2q_2}{n_2}} < \hat{p}_1-\hat{p}_2-(p_1-p_2) < z_{\frac{\alpha}{2}}\sqrt{\frac{p_1q_1}{n_1}+\frac{p_2q_2}{n_2}}\right] = 1-\alpha$$

Hence, for large samples the confidence interval for $p_1 - p_2$ is

$$(\hat{p}_1-\hat{p}_2)-z_{\frac{\alpha}{2}}\sqrt{\frac{p_1q_1}{n_1}+\frac{p_2q_2}{n_2}} < (p_1-p_2) < (\hat{p}_1-\hat{p}_2)+z_{\frac{\alpha}{2}}\sqrt{\frac{p_1q_1}{n_1}+\frac{p_2q_2}{n_2}}$$

where $\hat{p}_1, \hat{p}_2, \hat{q}_1,$ and $\hat{q}_2$ are the estimates of p_1, p_2, q_1, and q_2 given by $\hat{p}_1 = \frac{x_1}{n_1}$, $\hat{p}_2 = \frac{x_2}{n_2}$, $\hat{q}_1 = 1-\hat{p}_1$, $\hat{q}_2 = 1-\hat{p}_2$, and $z_{\frac{\alpha}{2}}$ is the value of z leaving an area of $\frac{\alpha}{2}$ to the right.

Worked Out Examples

EXAMPLE 9.1

Suppose that the heights of 100 male students at an Indian university represent a random sample of the heights of all 1,546 male students at the university. The following table shows the heights of 100 male students:

Height in inches	No. of students
60–62	5
63–65	18
66–68	42
69–71	27
72–74	8
Total	100

(i) Determine unbiased and efficient estimate of the true mean and true variance.

(ii) Find 95% confidence interval for estimating the mean height of students of the university.

Solution:

(i) To find the sample mean:

Height (inches)	Mid value (x_i)	Frequency (f_i)	f_ix_i	$(x-\bar{x})$ $\bar{x} = 67.45$	$F(x-\bar{x})^2$
60–62	61	5	305	–6.45	208.0125
63–65	64	18	1152	–3.45	214.2450
66–68	65	42	2814	–0.45	8.5050
69–71	70	27	1890	2.55	157.5675
72–74	73	8	584	5.55	246.42
		$\Sigma f_i = 100$		$\Sigma f_ix_i = 6745$	$\Sigma f(x-\bar{x})^2 = 852.75$

Mean of the above distribution $\bar{x} = \frac{\Sigma f_i x_i}{\Sigma f_i}$

$$= \frac{6745}{100}$$
$$= 67.45$$

Hence, unbiased and efficient estimate of the true mean height is $\bar{x} = 67.45$ inches.

Sample variance $S = \sqrt{\frac{\Sigma f (x - \bar{x})^2}{n}}$

$$= \sqrt{\frac{852.75}{100}}$$

$$= 2.92 \text{ inches}$$

(ii) The confidence interval for estimating the mean height of the students of the university is $\left(\bar{x} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right)$.

Since σ is not known, it can be replaced with S. When $\alpha = 95\%$, $z_{\frac{\alpha}{2}} = 1.96$ the confidence interval is

$$= \left[67.45 - (1.96)\left(\frac{2.92}{\sqrt{100}} \right), 67.45 + (1.96)\left(\frac{2.92}{\sqrt{100}} \right) \right]$$

$$= (66.88, 68.02)$$

(i.e.,) we can say with 100% confidence that the true mean (population mean) lies in the interval (66.88, 68.02).

EXAMPLE 9.2

The mean diameter of a random sample of 200 ball bearings made by a certain machine during one week is 0.824 inches and standard deviation is 0.042 inches. Find

(i) 95%

(ii) 99% confidence interval for the mean diameter of all the ball bearings.

Solution: Given mean of the sample, $\bar{x} = 0.824$

SD of the sample, $S = 0.042$
Size of the sample, $n = 200$

(i) 95% confidence interval for estimating the mean diameter of all ball bearings is

$$= \left[\bar{x} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right]$$

$$= \left[0.824 - (1.96)\left(\frac{0.042}{\sqrt{200}}\right), 0.824 + (1.96)\left(\frac{0.042}{\sqrt{200}}\right)\right]$$

$$= [0.824 - (1.96)(0.00296), 0.824 + (1.96)(0.00296)]$$

$$= (0.818, 0.829)$$

(ii) 99% confidence interval for estimating the mean diameter is

$$= \left[\bar{x} - z_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}, \bar{x} + z_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}\right]$$

where $1 - \alpha = 0.99$, $\alpha = 0.01$, $\frac{\alpha}{2} = 0.005$, $z_{\frac{\alpha}{2}} = 2.575$

$$= \left[0.824 - 2.575\left(\frac{0.042}{\sqrt{200}}\right), 0.824 + 2.575\left(\frac{0.042}{\sqrt{200}}\right)\right]$$

$$= (0.824 - 0.00764, 0.824 + 0.00764)$$

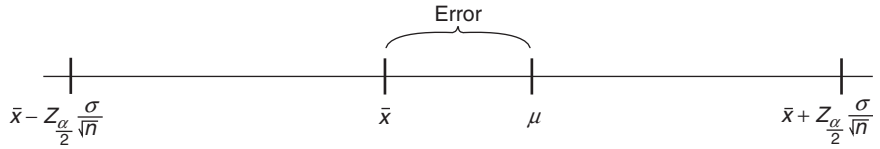
$$= (0.81635, 0.8316)$$

9.5 SOME RESULTS

Let us now find some formulae to find the maximum error of estimate and the sample size using the theory of point estimation. We observed that $(1 - \alpha)100\%$ confidence interval provides an estimate of the accuracy to a point estimate. If we consider the confidence interval for μ,

$$\bar{x} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

μ is the central value of the interval and $\bar{x}$ is not an estimate of it. Most of the time $\bar{x}$ will not equal μ and there is an error in the point estimate. The size of this error is the absolute value of the difference between μ and $\bar{x}$ and we can assert with $100(1 - \alpha)\%$ confidence that this difference will not exceed $z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$.



Error in estimating μ by $\bar{x}$

Result 1: If $\bar{x}$ is used as estimate of μ, we can assert with $100(1 - \alpha)\%$ confidence that the error will not exceed $z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$.

Hence, maximum error of estimate is given by $e = z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$.

Result 2: If $\bar{x}$ is used as an estimate of μ, we can be $(1 - \alpha)$ 100 % confident that the error will not exceed a specified amount e (given in the previous result) when the sample size is

$$e = z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

$$\sqrt{n} = \frac{\sigma z_{\frac{\alpha}{2}}}{e}$$

$$n = \left(\frac{z_{\frac{\alpha}{2}} \sigma}{e} \right)^2$$

Caution:

- The value obtained using result 2 has to rounded off to the nearest integer so that the degree of confidence never falls below 100 $(1 - \alpha)$ %.
- If we know the variance of the population σ, then we can find the sample size. Even otherwise, to find an estimate of a sample namely S, use the sample size $n \geq 30$. Then using S, we can determine approximately how many observations are needed to provide the desired degree of accuracy.
- For small samples, maximum error of estimate $e = t_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$.

Worked Out Examples

EXAMPLE 9.3

What is the maximum error one can expect to make probability 0.95 when using the mean of a random sample of size $n = 100$ to estimate the mean of a population with $\sigma^2 = 4.41$?

Solution: Given the probability = 0.95

$1 - \alpha = 0.95$, the value of z is $z_{\frac{\alpha}{2}} = 1.96$

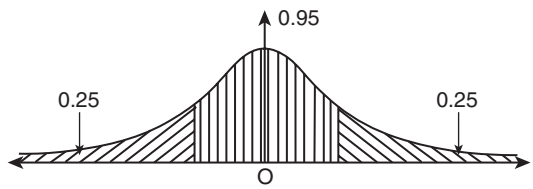
Given $\sigma^2 = 4.41$, $\sigma = 2.1$, and $n = 100$

Maximum error, $e = z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$

$$= \frac{(1.96)(2.1)}{\sqrt{100}}$$

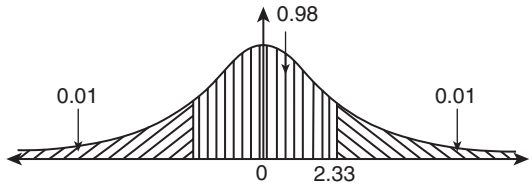
$$= \frac{4.116}{10}$$

$$= 0.4116$$



EXAMPLE 9.4

The principal of a college wants to use the mean of a random sample to estimate the average amount of time, students take to get from one class to the next, and he wants to be able to assert with 98% confidence that the error is at most 0.25 minutes. If it can be presumed from experience that $\sigma = 1.25$ minutes, how large a sample will he has to take?



Solution: Given that $1 - \alpha = 0.98$

$$\alpha = 0.02$$

$$z_{\frac{\alpha}{2}} = 2.33$$

Given that $\sigma = 1.25$ and maximum error $e = 0.25$

$$\therefore \text{ Sample size } n = \left(\frac{z_{\frac{\alpha}{2}} \sigma}{e} \right)^2$$

$$= \left[\frac{(2.33)(1.25)}{0.25} \right]^2$$

$$= 135.7225$$

$$\simeq 136$$

The sample size must be at most 136.

EXAMPLE 9.5

The mean muscular endurance score of a random sample of 60 subjects was found to be 145 with a SD of 40.

(i) Construct a 95% confidence interval for the true mean.

(ii) What size of the sample is required to estimate the mean within 5 of the true mean with 95% confidence?

Solution:

(i) Given that $n = 60$,

Mean of the sample $\bar{x} = 145$

SD of the sample $S = 40$

$1 - \alpha = 0.95$

$\alpha = 0.05$

$z_{\frac{\alpha}{2}} = 1.96$

The 95% confidence limits for true mean μ are $\left(\bar{x} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}},\ \bar{x} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right)$

$$= \left[145 - \frac{(1.96)(40)}{\sqrt{60}},\ 145 + \frac{(1.96)(40)}{\sqrt{60}} \right]$$

$$= \left(145 \pm \frac{78.4}{7.75}\right) = 145 \pm 10.12$$

$$= (134.88, 155.12)$$

(ii) The maximum sample of the estimate of mean is $n = \left(\frac{z_{\frac{\alpha}{2}}\sigma}{e}\right)^2$

where maximum error of estimate, $e = 5$

$$n = \left[\frac{(1.96)(40)}{5}\right]^2 = (15.68)^2$$

$$= 245.86$$

$$n \simeq 246$$

{we have the estimate of σ given by $S = 40$ and the absolute difference is given $|\bar{x} - \mu| \leq 5 = e$}.

EXAMPLE 9.6

Find 95% confidence limits for the mean of a normally distributed population from which the following sample was taken: 15, 17, 10, 18, 16, 9, 7, 11, 13, 14.

Solution: Mean of the sample $\bar{x} = \frac{[15+17+10+18+16+9+7+11+13+14]}{10}$

$$= \frac{130}{10}$$

$$= 13$$

The standard deviation of the sample is

$$S^2 = \frac{1}{n-1}\Sigma(x_i = \bar{x})^2$$

$$= \frac{1}{9}\begin{bmatrix}(15-13)^2 + (17-13)^2 + (10-13)^2 + (18-13)^2 + (16-13)^2 \\ +(9-13)^2 + (7-13)^2 + (11-13)^2 + (13-13)^2 + (14-13)^2\end{bmatrix}$$

$$= \frac{1}{9}[4+16+9+25+9+16+36+4+0+1]$$

$$= \frac{120}{9} = 13.33$$

$$S = 3.65$$

Since the given sample is small, the confidence limits for μ are $\left(\bar{x} - t_{\frac{\alpha}{2}}\frac{S}{\sqrt{n}},\ \bar{x} + t_{\frac{\alpha}{2}}\frac{S}{\sqrt{n}}\right)$

where $1 - \frac{\alpha}{2} = 0.975$

$$t_{\frac{\alpha}{2}} = 2.26$$

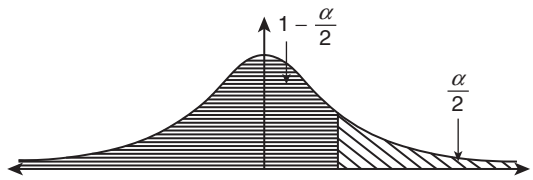
for $n - 1 = 10 - 1 = 9$ degrees of freedom.

Hence the confidence limits are

$$= \left[13 - \frac{2.26(3.65)}{\sqrt{10}}, 13 + \frac{2.26(3.65)}{\sqrt{10}}\right]$$

$$= \left[13 - \frac{8.249}{\sqrt{10}}, 13 + \frac{8.249}{\sqrt{10}}\right]$$

$$= (10.4, 15.6)$$



EXAMPLE 9.7

A random sample of 100 teachers in a large metropolitan area revealed a mean weekly salary of ₹2,000 with a standard deviation of ₹43. With what degree of confidence can we assert that the average weekly salary of all teachers in the metropolitan area is between ₹1,985 and ₹2,015?

Solution: Given mean weakly sales $\mu = 2000$

Standard deviation $\sigma = 43$

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{\bar{x} - 2000}{\frac{43}{\sqrt{100}}} = \frac{\bar{x} - 2000}{4.3}$$

The standard variable corresponding to ₹1,985

$$z_1 = \frac{1985 - 2000}{4.3}$$

$$= -3.4883$$

The standard variable corresponding to ₹2,015

$$z_2 = \frac{2015 - 2000}{4.3}$$

$$= 3.4883$$

If X is the mean weekly salary then

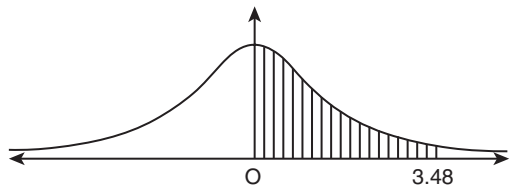
$$P(1985 < X < 2015) = P(-3.48 < Z < 3.48)$$

$$= 2P(0 < Z < 3.48)$$

$$= 2(0.4997)$$

$$= 0.9994$$

Thus, we can ascertain with 99.94% confidence that average weekly salary of all teachers in the metropolitan area is between ₹1,985 and ₹2,015.



EXAMPLE 9.8

A random sample of size 10 from a central was taken with standard deviation 0.03. Find the maximum error with 99% confidence.

Solution: Given sample size $n = 10$

Sample standard deviation $S = 0.03$

$$1 - \frac{\alpha}{2} = 0.995 \quad \alpha = 0.01$$

$$\frac{\alpha}{2} = 0.005$$

$t_{\frac{\alpha}{2}}$ for $n - 1 = 10 - 1 = 9$ degrees of freedom

$$t_{\frac{\alpha}{2}} = 3.25$$

$\therefore$ Maximum error of estimate $e = t_{\frac{\alpha}{2}} \frac{S}{\sqrt{n}}$

$$= \frac{(3.25)(0.03)}{\sqrt{10}}$$

$$= 0.0325$$

Result 3: The maximum error of estimate of proportion p is given by

$$e = z_{\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{n}}$$

Here the magnitude of the error we make when we use $\frac{X}{n}$ as an estimate for p is $\left|\left(\frac{X}{n} - p\right)\right|$. We can assert with probability $(1 - \alpha)$ that

$$\left|\frac{X}{n} - p\right| \leq z_{\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{n}}$$

Result 4: Hence, the *maximum sample size* that is needed to estimate proportion p is

$$e = z_{\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{n}}$$

$$\sqrt{n} = \frac{z_{\frac{\alpha}{2}}}{e} \sqrt{p(1-p)}$$

$$n = p(1-p)\left(\frac{z_{\frac{\alpha}{2}}}{e}\right)^2$$

Result 5: If the proportion p is unknown, then we can use the fact that $p(1 - p) = \frac{1}{4}$

$$\text{(i.e.,) } p = \frac{1}{2}$$

The sample size for the above case is

$$n = \frac{1}{4}\left(\frac{z_{\frac{\alpha}{2}}}{e}\right)^2$$

EXAMPLE 9.9

Out of 3,470 recently painted grey automobiles, 468 had paint problems that could easily be detected by visual inspection. Obtain a 95% confidence interval for the population proportion of defective grey paint jobs.

Solution: The confidence interval for proportion p is

$$\left(\hat{p} - z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}\hat{q}}{n}}, \hat{P} + Z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}\hat{q}}{n}}\right)$$

where $\hat{p} = \frac{X}{n}$ and $\hat{q} = 1 - \frac{X}{n}$

Here X = number of automobiles painted recently that have problems
= 468
n = Total number of automobiles painted recently
= 3470

$$\hat{p} = \frac{X}{n} = \frac{468}{3470} = 0.1348 \sim 0.135$$

$$\hat{q} = 0.865$$

Given $1 - \alpha = 0.95$

$$z_{\frac{\alpha}{2}} = 1.96$$

$\therefore$ The required confidence interval is

$$= \left[(0.135) - (1.96)\sqrt{\frac{(0.135)(0.865)}{3470}}, 0.135 + 1.96\sqrt{\frac{(0.135)(0.865)}{3470}}\right]$$

$$= (0.135 - 0.01136, 0.135 + 0.01136)$$

$$= (0.1236, 0.1463)$$

Hence, the confidence interval for proportion p is (0.123, 0.146)

EXAMPLE 9.10

In a sample survey conducted in a large city 136 out of 400 persons answered yes to the question of whether their city's public transportation is adequate. What can you say about the maximum error with 99% confidence?

Solution: Given X = no. of persons who answered yes
= 136
n = number of persons in survey
= 400

$$p = \frac{X}{n} = 0.34$$

$$1 - p = 0.66$$

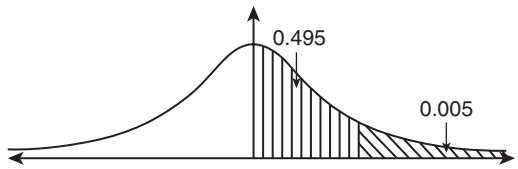
Maximum error of estimate $e = z_{\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}}$

Given $1-\alpha = 0.99$

$\alpha = 0.01$

$\frac{\alpha}{2} = 0.005$

$z_{\frac{\alpha}{2}} = 2.58$



$\therefore e = 2.58\sqrt{\frac{(0.34)(0.66)}{400}} = 0.061$

EXAMPLE 9.11

In a sample survey of the safety explosives used in certain mining operations, explosives containing potassium nitrate were found to be used in 120 out of 300 cases.

(i) Construct 95% confidence interval for corresponding true proportion.

(ii) What is the maximum error obtained in estimating $\frac{X}{n}$ as the true proportion?

Solution:

(i) Given X = explosives containing potassium nitrate = 120

n = explosives used in mining operations = 300

Proportion, $p = \frac{X}{n} = \frac{120}{300} = 0.4$

$1 - p = 0.6$

Confidence interval for true proportion is

$$= \left[\hat{p} - Z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}\,\hat{q}}{n}}, \hat{p} + Z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}\,\hat{q}}{n}}\right]$$

$$1-\alpha = 0.95,\ \alpha = 0.05,\ \frac{\alpha}{2} = 0.0025,\ Z_{\frac{\alpha}{2}} = 1.96$$

$$= \left(0.4 - (1.96)\sqrt{\frac{(0.4)(0.6)}{300}}, 0.4 + (1.96)\sqrt{\frac{(0.4)(0.6)}{300}}\right)$$

$$= (0.4 - 0.0554,\ 0.4 + 0.0554)$$

$$= (0.3445,\ 0.4554)$$

Hence 95% confidence interval for true proportion is (0.34, 0.45)

(ii) Maximum error of estimate

$$e = Z_{\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}}$$

Given $z_{\frac{\alpha}{2}} = 1.96$

$p = 0.4,\ 1 - p = 0.6$

Sample size, $n = 300$

$$\therefore \text{ Maximum error } e = 1.96\sqrt{\frac{(0.4)(0.6)}{300}}$$

$$= 1.96\ (0.0282)$$

$$= 0.0554$$

EXAMPLE 9.12

A geneticist is interested in the proportion of males who have a certain blood disorder. In a random sample of 100 males, 24 are found to be afflicted.

(i) Compute 99% confidence interval for the proportion of males who have this blood disorder.

(ii) What can we assert with 99% confidence about the possible size of our error if we estimate the proportion of males with blood disorder to be 0.24?

Solution:

(i) Given, size of the sample of males, $n = 100$

Males who have blood disorder, $X = 24$

Proportion of males with blood disorder, $p = \frac{X}{n} = \frac{24}{100} = 0.24$

$$q = 1 - p = 0.76$$

$$1 - \alpha = 0.99,\ \alpha = 0.01,\ \frac{\alpha}{2} = 0.005$$

$$z_{\frac{\alpha}{2}} = 2.575$$

Confidence interval for true proportion is $\left[\hat{p} - Z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}\hat{q}}{n}},\ \hat{p} + Z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}\hat{q}}{n}}\right]$

$$= \left[0.24 - 2.575\sqrt{\frac{(0.24)(0.76)}{100}},\ 0.24 + 2.575\sqrt{\frac{(0.24)(0.76)}{100}}\right]$$

$$= (0.24 - 0.1099,\ 0.24 + 0.1099)$$

$$\simeq (0.13,\ 0.349)$$

(ii) The maximum error of estimate is given by

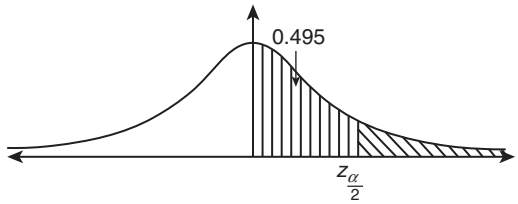
$$e = z_{\frac{\alpha}{2}}\frac{\sqrt{p(1-p)}}{\sqrt{n}}$$

$$\text{Size of the error } n = \left[Z_{\frac{\alpha}{2}}\frac{p(1-p)}{e}\right]^2$$

$$\text{Possible size of the error } e = 2.575\sqrt{\frac{(0.24)(0.76)}{100}}$$

$$= 0.1099$$

$$1-\alpha = 0.99$$
$$\alpha = 0.01$$
$$\frac{\alpha}{2} = 0.005$$
$$z_{\frac{\alpha}{2}} = 2.575$$



EXAMPLE 9.13

A clinical trial is conducted to determine if a certain type of inoculation has an effect on the incidence of a certain disease. A sample of 1,000 rats was kept in a controlled environment for a period of 1 year and 500 of the rats were given the inoculation. Of the group not given the drug, there were 120 incidences of the disease, while 98 of the inoculated group contracted it. Construct 90% confidence interval for $p_1 - p_2$ where p_1 is probability of incidence of the disease in uninoculated rats and p_2 is the probability of incidence after receiving the drug.

Solution: Given data:

Let X_1 = rats which are given inoculation and with disease = 98
n_1 = sample of rats which are given inoculation = 500
X_2 = rats with disease and not inoculated = 120
n_2 = sample of rats not inoculated = 500

$$p_1 = \frac{X_1}{n_1} = \frac{98}{500} = 0.196 \qquad 1-\alpha = 0.90$$
$$q_1 = 1 - p_1 = 0.804 \qquad \alpha = 0.1$$
$$p_2 = \frac{X_2}{n_2} = \frac{120}{500} = 0.24 \qquad \frac{\alpha}{2} = 0.05$$
$$q_2 = 1 - p_2 = 0.76 \qquad z_{\frac{\alpha}{2}} = 1.645$$

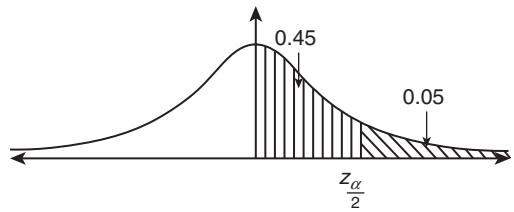
The confidence interval for $p_2 - p_1$ is

$$\left[(p_2 - p_1) - z_{\frac{\alpha}{2}}\sqrt{\frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}}, (p_2 - p_1) + z_{\frac{\alpha}{2}}\sqrt{\frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}}\right]$$

$$\left[(0.24 - 0.196) \pm 1.645\sqrt{\frac{(0.196)(0.804) + (0.24)(0.76)}{500}}\right]$$

$$\left[0.044 \pm (1.645)\sqrt{\frac{(0.1575) + (0.1824)}{500}}\right]$$

$$(0.044 \pm 0.04288) = (0.044 - 0.14288, 0.044 + 0.04288)$$
$$= (0.0011, 0.0868)$$



EXAMPLE 9.14

Assuming that $\sigma = 20.0$, how large a random sample be taken to assert with probability 0.95 that the sample mean will not differ from true mean by more than 3.0.

Solution: Given that

Maximum error $E = |\bar{x} - \mu| = 3$

Standard deviation $\sigma = 20.0$

Probability, $1 - \alpha = 0.95$, hence $z_{\frac{\alpha}{2}} = 1.96$

Hence the sample size, $n = \left(\frac{z_{\frac{\alpha}{2}} \sigma}{E} \right)^2 = \left[\frac{1.96(20)}{3} \right]^2$

$$n = 170.7377 \cong 171$$

EXAMPLE 9.15

400 articles in a factory were examined and 3% were found to be defective. Construct 95% confidence interval for proportion of defective.

Solution: Given that sample size of articles in a factory, $n = 400$

Let X be the number of defective articles = 3% = $\frac{3}{100} \times 400$

$\therefore X = 12$

Let p = proportion of defective articles in the sample

$$= \frac{X}{n} = \frac{12}{400} = 0.03$$

$$q = 1 - p = 1 - 0.03 = 0.97$$

In addition, given that probability $1 - \alpha = 0.95$

$$\therefore z_{\frac{\alpha}{2}} = 1.96$$

Confidence interval for proportion p is $\hat{p} - z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}\hat{q}}{n}},\ \hat{p} + z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}\hat{q}}{n}}$

$$= \left[0.03 - 1.96 \sqrt{\frac{(0.03)(0.97)}{400}},\ 0.03 + 1.96 \sqrt{\frac{(0.03)(0.97)}{400}} \right]$$

$$= (0.03 - 0.0167, 0.03 + 0.0167)$$

$$= (0.01328, 0.0467)$$

Hence, confidence interval is (0.0133, 0.0467)

EXAMPLE 9.16

In a study of an automobile insurance a random sample of 80 body repair costs had a mean of ₹472.36 with what confidence can we assert that maximum error does not exceed ₹10, when the standard deviation is ₹62.35.

Solution: Given:

Maximum error $E = ₹10$

Standard deviation, $\sigma = 62.35$

Sample size, $n = 80$

We know that $$E = z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

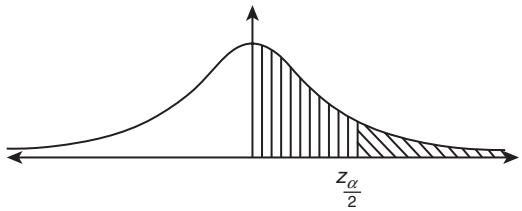
$$10 = z_{\frac{\alpha}{2}} \frac{62.35}{\sqrt{80}}$$

$$z_{\frac{\alpha}{2}} = \frac{10\sqrt{80}}{62.35} = 1.43$$

$$1 - \alpha = 2$$

We got $z_{\frac{\alpha}{2}} = 1.43$

From tables, from the graph $P(0 < z < 1.43) = 0.4236$



i.e., $\frac{\alpha}{2} = 0.4236, \qquad \alpha = 0.8472$

Hence, we assert with 84.72% that maximum error will not exceed ₹10.

EXAMPLE 9.17

If 80 patients are treated with an antibiotic 59 got cured. Find 95% confidence limits for true proportion of cure.

Solution: Number of patients, $n = 80$

Patients treated with antibiotic, $X = 59$

Proportion of cured patients, $p = \frac{X}{n} = \frac{59}{80} = 0.7375$

$1 - \alpha = 0.95$, $\alpha = 0.05$, $\frac{\alpha}{2} = 0.025$, $z_{\frac{\alpha}{2}} = 1.96$, $q = 1 - p = 0.2625$

Confidence limit for true proportion is $\left[\hat{p} - z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}\hat{q}}{n}},\ \hat{p} + z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}\hat{q}}{n}}\right]$

$$= \left[0.7375 - 1.96\sqrt{\frac{(0.7375)(0.2625)}{80}},\ 0.7375 + 1.96\sqrt{\frac{(0.7375)(0.2625)}{80}}\right]$$

$$= [0.7375 - (1.96)(0.04919), 0.7375 + (1.96)(0.04919)]$$

$$= (0.7375 - 0.0964, 0.7375 + 0.0964)$$

$$= (0.64108, 0.8339)$$

Hence, confidence limit for true proportion is (0.64, 0.83).

9.6 CONFIDENCE INTERVAL FOR DIFFERENCE BETWEEN TWO MEANS (KNOWN VARIANCES)

Now let us derive the formula for confidence interval between two means by considering two cases when the variances are known and unknown and also two cases for large and small samples.

Let us select two independent random samples of sizes n_1 and n_2 from two different populations. We compute their sample means $\bar{x}_1$ and $\bar{x}_2$ and their difference is given as $\bar{x}_1 - \bar{x}_2$. Then the sampling distribution of $\bar{x}_1 - \bar{x}_2$ is approximately normally distributed with mean.

$\mu_{\bar{x}_1 - \bar{x}_2} = \mu_1 - \mu_2$ and standard deviation $\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$.

Hence, we can assert with $100(1 - \alpha)$ % confidence that

$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_1^2}{n_2}}}$ is a standard normal variate which falls between $-z_{\frac{\alpha}{2}}$ and $z_{\frac{\alpha}{2}}$.

$$\text{(i.e.,)} \quad P(-z_{\frac{\alpha}{2}} < z < z_{\frac{\alpha}{2}}) = 1 - \alpha$$

$$P\left[-z_{\frac{\alpha}{2}} < \frac{(X_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_1^2}{n_2}}} < z_{\frac{\alpha}{2}}\right] = 1 - \alpha$$

Hence, the $100(1 - \alpha)$% confidence interval for $(\mu_1 - \mu_2)$ where $\bar{x}_1$ and $\bar{x}_2$ are the means of two independent random samples of sizes n_1 and n_2 with known variances σ_1^2 and σ_2^2 is

$$(\bar{x}_1 - \bar{x}_2) - z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_1^2}{n_2}} < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_1^2}{n_2}}$$

where $z_{\frac{\alpha}{2}}$ is the z-value leaving area of $\frac{\alpha}{2}$ to the right.

9.7 CONFIDENCE INTERVAL FOR DIFFERENCE BETWEEN TWO MEANS (UNKNOWN VARIANCES)

When we consider that the variances are unknown, their estimates can be taken. Let us consider two independent samples of sizes n_1 and n_2 be drawn from two normal populations with means μ_1 and μ_2. The pooled estimate of the unknown common variance σ^2 can be obtained by pooling the sample variances, $S_1^2$and S_2^2.

$$S^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

Then the sampling distribution of $(\mu_1 - \mu_2)$ has the t-distribution with $(n_1 + n_2 - 2)$ degrees of freedom and we can assert with $100(1 - \alpha)$% confidence that $(\mu_1 - \mu_2)$ lies between $-t_{\frac{\alpha}{2}}$ and $t_{\frac{\alpha}{2}}$.

$$\text{(i.e.,)} \quad P(-t_{\frac{\alpha}{2}} < t < t_{\frac{\alpha}{2}}) = 1 - \alpha$$

$$P\left[-t_{\frac{\alpha}{2}} < \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} < t_{\frac{\alpha}{2}}\right] = 1 - \alpha$$

Hence, if $\bar{x}_1$ and $\bar{x}_2$ are the sample means of two independent random samples of sizes n_1 and n_2 drawn from normal populations with the unknown but equal variances, then a $100(1 - \alpha)\%$ confidence interval for $(\mu_1 - \mu_2)$ is given by

$$(\bar{x}_1 - \bar{x}_2) - t_{\frac{\alpha}{2}} S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} < (\mu_1 - \mu_2) < (\bar{x}_1 - \bar{x}_2) + t_{\frac{\alpha}{2}} S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

where S is the pooled estimate of the population standard deviation and $t_{\frac{\alpha}{2}}$ is the value of t with $(n_1 + n_2 - 2)$ degrees of freedom, leaving an area of $\frac{\alpha}{2}$ to the right.

9.8 CONFIDENCE INTERVAL FOR DIFFERENCE OF MEANS (UNKNOWN AND UNEQUAL VARIANCES)

Let $\bar{x}_1$ and $\bar{x}_2$ be the sample means and the two samples drawn from two normal populations with unknown and unequal variances. An approximate $100(1 - \alpha)\%$ confidence interval for $\mu_1 - \mu_2$ is obtained as follows:

$$P(-t_{\frac{\alpha}{2}} < t < t_{\frac{\alpha}{2}}) = 1 - \alpha$$

$$P\left[-t_{\frac{\alpha}{2}} < \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}} < t_{\frac{\alpha}{2}}\right] = 1 - \alpha$$

$$(\bar{x}_1 - \bar{x}_2) - t_{\frac{\alpha}{2}}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} < (\mu_1 - \mu_2) < (\bar{x}_1 - \bar{x}_2) + t_{\frac{\alpha}{2}}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

where $t_{\frac{\alpha}{2}}$ is the value of t with $(n_1 + n_2 - 2)$ degrees of freedom, leaving an area of $\frac{\alpha}{2}$ to the right.

9.9 CONFIDENCE INTERVAL FOR DIFFERENCE BETWEEN MEANS FOR PAIRED OBSERVATIONS

In the earlier topics, we have seen the estimation procedures for difference between two means when variances are known and unknown. Now let us deal with a special case of paired observations. The conditions of the two populations are not assigned randomly to experimental units. Rather each homogeneous experimental units receives both population conditions and as a result, each experimental unit has a pair of observations one for each population.

Let the pair be represented by (x_i, y_i) whose differences are denoted by $d_i = x_i - y_i$ for $i = 1, 2, \ldots n$ observations of a sample. Let the set of observations be drawn from normal population with mean $\mu_d = \mu_1 - \mu_2$ and variance σ_d^2, whose estimate is S_d^2.

A $(1 - \alpha)100\%$ confidence interval for μ_d whose estimator is $\bar{d}$ can be obtained as

$P(-t_{\frac{\alpha}{2}} < t < t_{\frac{\alpha}{2}}) = 1-\alpha$ where $t = \dfrac{\bar{d}-\mu_d}{\frac{S_d}{\sqrt{n}}}$ follows t-distribution with $(n-1)$ degrees of freedom and $t_{\frac{\alpha}{2}}$ is the value of t leaving an area of $\dfrac{\alpha}{2}$ to the right,

$$\mu_d = \mu_1 - \mu_2 \qquad \bar{d} = \frac{\sum_{i=1}^{n} d_i}{n}$$

$$S_d = \sqrt{\frac{\sum (di-\bar{d})^2}{n-1}}$$

Hence a $(1-\alpha)$ 100% confidence interval for $\mu_d = \mu_1 - \mu_2$ where $\bar{d}$ and S_d are the mean and standard deviation of the normally distributed differences of n random pairs of measurements, is

$$P\left[-t_{\frac{\alpha}{2}} < \frac{\bar{d}-\mu_d}{\frac{S_d}{\sqrt{n}}} < t_{\frac{\alpha}{2}}\right] = 1-\alpha$$

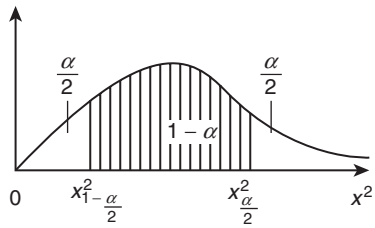
(i.e.,)
$$\bar{d} - t_{\frac{\alpha}{2}}\frac{S_d}{\sqrt{n}} < \mu_d < \bar{d} + t_{\frac{\alpha}{2}}\frac{S_d}{\sqrt{n}}$$

9.10 CONFIDENCE INTERVAL FOR ESTIMATING THE VARIANCE

Let a sample of size n be drawn from a normal population whose variance is σ^2. Let the sample variance S^2 be an estimator of σ^2. An interval estimate of σ^2 can be obtained using the statistic $\chi^2 = \dfrac{(n-1)S^2}{\sigma^2}$ which follows χ^2 distribution with $(n-1)$ degree of freedom.

Hence, $P(\chi^2_{1-\frac{\alpha}{2}} < \chi^2 < \chi^2_{\frac{\alpha}{2}}) = 1-\alpha$ which is shown as follows:

Here $\chi^2_{1-\frac{\alpha}{2}}$ and χ^2_{α} are the values of χ^2 distribution with $(n-1)$ degrees of freedom leaving areas of $1-\dfrac{\alpha}{2}$ and $\dfrac{\alpha}{2}$ to the right, respectively.



$$P\left[\chi^2_{1-\frac{\alpha}{2}} < \frac{(n-1)S^2}{\sigma^2} < \chi^2_{\frac{\alpha}{2}}\right] = 1-\alpha$$

$$P\left[\frac{\chi^2_{1-\frac{\alpha}{2}}}{(n-1)S^2} < \frac{1}{\sigma^2} < \frac{\chi^2_{\frac{\alpha}{2}}}{(n-1)S^2}\right] = 1-\alpha$$

$$P\left[\frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2}}} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2}}}\right] = 1-\alpha$$

Hence a $(1-\alpha)100\%$ confidence interval for σ^2 where S^2 is the variance of a random sample of size n drawn from a normal population is

$$\frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2}}} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2}}}$$

where $\chi^2_{\frac{\alpha}{2}}$ and $\chi^2_{1-\frac{\alpha}{2}}$ are the values of χ^2 with $(n-1)$ degrees of freedom leaving areas $\frac{\alpha}{2}$ and $1-\frac{\alpha}{2}$ to the right, respectively.

9.11 CONFIDENCE INTERVAL FOR ESTIMATING THE RATIO OF TWO VARIANCES

Let two random samples of sizes n_1 and n_2 with sample variances S_1^2 and S_2^2 be drawn from normal populations with variances σ_1^2 and σ_2^2. Let $\frac{S_1^2}{S_2^2}$ be the estimator of $\frac{\sigma_1^2}{\sigma_2^2}$. An interval estimate of the $\frac{\sigma_1^2}{\sigma_2^2}$ can be obtained using the statistic

$$F = \frac{\left(\frac{S_1^2}{\sigma_1^2}\right)}{\left(\frac{S_2^2}{\sigma_2^2}\right)}$$

(i.e.,) $F = \frac{\sigma_2^2 s_1^2}{\sigma_1^2 s_2^2}$, which follows F-distribution with $(n_1 - 1, n_2 - 1)$ degrees of freedom given by (γ_1, γ_2).

Hence, $P[F_{1-\frac{\alpha}{2}}(\gamma_1, \gamma_2) < F < F_{\frac{\alpha}{2}}(\gamma_1, \gamma_2)] = 1 - \alpha$, where $F_{1-\frac{\alpha}{2}}(\gamma_1, \gamma_2)$ and $F_{\frac{\alpha}{2}}(\gamma_1, \gamma_2)$ are the values of F-distribution with $(\gamma_1, \gamma_2) = (n_1 - 1, n_2 - 1)$ degrees of freedom, leaving areas $\left(1-\frac{\alpha}{2}\right)$ and $\frac{\alpha}{2}$ to the right, respectively as shown in the following figure:

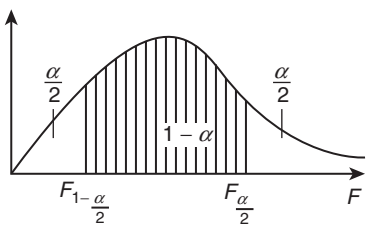
Hence,

$$P\left[F_{1-\frac{\alpha}{2}}(\gamma_1, \gamma_2) < \frac{\sigma_2^2 s_1^2}{\sigma_1^2 s_2^2} < F_{\frac{\alpha}{2}}(\gamma_1, \gamma_2)\right] = 1-\alpha$$

$$P\left[\frac{F_{1-\frac{\alpha}{2}}(\gamma_1, \gamma_2)}{s_1^2/s_2^2} < \frac{\sigma_2^2}{\sigma_1^2} < \frac{F_{\frac{\alpha}{2}}(\gamma_1, \gamma_2)}{s_1^2/s_2^2}\right] = 1-\alpha$$

$$P\left[\frac{s_2^2}{s_1^2} F_{1-\frac{\alpha}{2}}(\gamma_1, \gamma_2) < \frac{\sigma_2^2}{\sigma_1^2} < \frac{s_2^2}{s_1^2} F_{\frac{\alpha}{2}}(\gamma_1, \gamma_2)\right] = 1-\alpha$$

$$P\left[\frac{s_1^2}{s_2^2 F_{\frac{\alpha}{2}}(\gamma_1, \gamma_2)} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{s_1^2}{s_2^2 F_{1-\frac{\alpha}{2}}(\gamma_1, \gamma_2)}\right] = 1-\alpha$$



Hence, a $(1-\alpha)100\%$ confidence interval for $\frac{\sigma_1^2}{\sigma_2^2}$ where s_1^2 and s_2^2 are the variances of independent samples of sizes n_1 and n_2 which are drawn from normal populations is

$$\frac{s_1^2}{s_2^2 F_{\frac{\alpha}{2}}(\gamma_1, \gamma_2)} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{s_1^2}{s_2^2 F_{1-\frac{\alpha}{2}}(\gamma_1, \gamma_2)}$$

where $F_{\frac{\alpha}{2}}(\gamma_1, \gamma_2)$ and $F_{1-\frac{\alpha}{2}}(\gamma_1, \gamma_2)$ are the value of F-distribution with $(\gamma_1, \gamma_2) = (n_1 - 1, n_2 - 1)$ degrees of freedom leaving areas $\frac{\alpha}{2}, 1-\frac{\alpha}{2}$ to the right, respectively.

Worded Out Examples

EXAMPLE 9.18

The scores on a standardized math test in a district X are normally distributed with mean 74 and standard deviation 8 and size 40, while those in district Y are normally distributed with mean 70 and standard deviation 10 and size 50. Find 95% confidence interval for difference of their averages.

Solution: Given the data of district X as:

Sample mean $\bar{x}_1 = 74$,

Sample size $n_1 = 40$

Sample standard deviation $s_1 = 8$

This is to be taken as an estimator of the population variance.

Given the data of district Y as:

Sample mean $\bar{x}_2 = 70$,

Sample size $n_2 = 50$

Sample standard deviation, $S_2 = 10$ (This is an estimator for population variance), $1 - \alpha = 0.95$, $\therefore\ z_{\frac{\alpha}{2}} = 1.96$. The $(1 - \alpha)100\%$ confidence interval for $(\mu_1 - \mu_2)$ is

$$= \left[(\bar{x}_1 - \bar{x}_2) - z_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}},\ (\bar{x}_1 - \bar{x}_2) + z_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right]$$

$\sigma_1^2 \simeq S_1^2, \qquad \sigma_2^2 \simeq S_2^2$

$$= \left[(\bar{x}_1 - \bar{x}_2) - z_{\frac{\alpha}{2}} \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}},\ (\bar{x}_1 - \bar{x}_2) + z_{\frac{\alpha}{2}} \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \right]$$

$$= \left[(74 - 70) - 1.96 \sqrt{\frac{8^2}{40} + \frac{10^2}{50}},\ (74 - 70) + 1.96 \sqrt{\frac{8^2}{40} + \frac{10^2}{50}} \right]$$

$$= [4 - 1.96(1.897),\ 4 + (1.96)(1.897)]$$

$$= [4 - 3.7188,\ 4 + 3.7188]$$

$$= (0.28,\ 7.7188) \simeq (0.28,\ 7.72)$$

EXAMPLE 9.19

Two kinds of thread are compared for strength. 50 pieces of each type of thread are tested under similar conditions. Brand I had an average tensile strength of 78.3 kilograms with a standard deviation of 5.6 kg while Brand II had an average tensile strength of 87.2 and standard deviation of 6.3 kg.

(i) Construct 95% confidence interval for difference of population means.

(ii) Construct 99% confidence interval for $(\mu_2 - \mu_1)$.

Solution: Given the data of

Brand I	Brand II
Sample mean, $\bar{x}_1 = 78.3$	Sample mean, $\bar{x}_2 = 87.2$
Sample standard deviation, $s_1 = 5.6$	Sample standard deviation, $s_2 = 6.3$
Sample size, $n_1 = 50$	Sample size, $n_2 = 50$

$$1 - \alpha = 95\%$$
$$z_{\frac{\alpha}{2}} = 1.96$$

The confidence interval for difference of population means $(\mu_2 - \mu_1)$ is

$$= \left[(\bar{x}_2 - \bar{x}_1) - z_{\frac{\alpha}{2}} \sqrt{\frac{s_2^2}{n_2} + \frac{s_1^2}{n_1}},\ (\bar{x}_2 - \bar{x}_1) + z_{\frac{\alpha}{2}} \sqrt{\frac{s_2^2}{n_2} + \frac{s_1^2}{n_1}} \right] \tag{9.3}$$

$$= \left[(87.2 - 78.3) - 1.96 \sqrt{\frac{5.6^2 + 6.3^2}{50}},\ (87.2 - 78.3) + 1.96 \sqrt{\frac{5.6^2 + 6.3^2}{50}} \right]$$

$$= [8.9 - 1.96(1.192),\ 8.9 + (1.96)(1.192)]$$

$$\simeq (6.56, 11.23)$$

(ii) $1 - \alpha = 0.99,\ \alpha = 0.01,\ \frac{\alpha}{2} = 0.005$

$$z_{\frac{\alpha}{2}} = 2.575$$

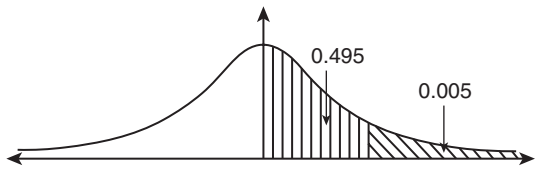
Substituting this in 9.3 we get,

$$= [(87.2 - 7.83) - (2.575)(1.192),\ (87.2 - 78.3) + (2.575)(1.192)]$$

$$\simeq [8.9 - (2.575)(1.192),\ 8.9 + (2.575)(1.192)]$$

$$= (5.8306, 11.9694)$$

Hence, 95% confidence interval for $(\mu_2 - \mu_1)$ is (6.56, 11.23) and 99% confidence interval for $(\mu_2 - \mu_1)$ is (5.83, 11.96).



EXAMPLE 9.20

Two catalysts are compared for their effect on the output of the process reaction in a batch of chemical process. The first sample of 12 batches was prepared using catalyst I gave an average yield of 85 with a standard deviation of 4 and the second sample of 10 batches was obtained using catalyst II which gave an average of 81 and sample standard deviation of 5. Find 90% confidence interval for the difference between population means assuming that the populations are approximately normally distributed with equal variances.

Solution: The given data is

Catalyst I	Catalyst II
Average yield, $\bar{x}_1 = 85$	Average yield, $\bar{x}_2 = 81$
Sample standard deviation, $S_1 = 4$	Sample standard deviation, $S_2 = 5$
Sample size, $n_1 = 12$	Sample size, $n_2 = 10$

Common pooled sample variance $S^2 = \dfrac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$

$$S^2 = \frac{11(4^2) + 9(5^2)}{12 + 10 - 2} = 20.05$$

$1 - \alpha = 0.90$, $\alpha = 0.01$, $\dfrac{\alpha}{2} = 0.05$ $\gamma = n_1 + n_2 - 2 = 20$

$t_{\frac{\alpha}{2}}$ (20 degrees of freedom) = 1.72

The confidence interval for $(\mu_1 - \mu_2)$ is given by,

$$= \left[(\bar{x}_1 - \bar{x}_2) - t_{\frac{\alpha}{2}} S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}},\ (\bar{x}_1 - \bar{x}_2) + t_{\frac{\alpha}{2}} S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right]$$

$$= \left[(85 - 81) - 1.72(4.47)\sqrt{0.183},\ (85 - 81) + 1.72(4.47)\sqrt{0.183} \right]$$

$$= [4 - 3.29,\ 4 + 3.29]$$

$$= (0.70,\ 7.29)$$

Hence, 90% confidence interval for $(\mu_1 - \mu_2)$ is (0.70, 7.29)

EXAMPLE 9.21

The following data recorded in days represents the length of time to recovery of patients randomly treated with one of the two medications to clear up bladder infections:

Medication 1	Medication 2
$n_1 = 14$	$n_1 = 16$
$\bar{x}_1 = 85$	$\bar{x}_2 = 19$
$S_1^2 = 1.5$	$S_2^2 = 1.8$

Find 99% confidence for the difference $\mu_2 - \mu_1$ in the mean recovery time for the two medications, assuming normal populations with equal variances.

Solution: The given data is as follows:

Common pooled sample variance is

$$S^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

$$= \frac{13(1.5) + 15(1.8)}{14 + 16 - 2}$$

$$S^2 = 1.6607$$

$$S = 1.288$$

The 99% confidence interval for $(\mu_2 - \mu_1)$ is given by,

$$\left[(x_2 - \bar{x}_1) - t_{\frac{\alpha}{2}} s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}},\ (x_2 - \bar{x}_1) + t_{\frac{\alpha}{2}} s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right]$$

where $t_{\frac{\alpha}{2}} = t_{0.005}$ at 28 degrees of freedom = 2.76

$$1 - \alpha = 0.99,\ \alpha = 0.01,\ \frac{\alpha}{2} = 0.005$$

$$\gamma = n_1 + n_2 - 2$$

$$= 28$$

$$\therefore \left[(19 - 17) - (2.76)(1.288)\sqrt{\frac{1}{14} + \frac{1}{16}},\ (19 - 17) + (2.76)(1.288)\sqrt{\frac{1}{14} + \frac{1}{16}}\right]$$

$$= [2 - (2.76)(1.288)(0.3659),\ 2 + (2.76)(1.288)(0.3659)]$$

$$= (0.6995,\ 3.3009)$$

Hence, 99% confidence interval for $\mu_2 - \mu_1$ is (0.69, 3.3).

EXAMPLE 9.22

The government awarded grants to the agricultural departments of 9 universities to test the yield capabilities of two new varieties of wheat. Each variety was planted on plots of equal area at each university and the yields in kilograms per plot are recorded as follows:

Varieties	Universities								
	1	2	3	4	5	6	7	8	9
1	38	23	35	41	44	29	37	31	38
2	45	25	31	38	50	33	36	40	43

Find 95% confidence interval for the mean difference between the yields of the two varieties assuming the differences of yields to be approximately normally distributed.

Solution: Since the same set of plots are considered for two different yields, pairing of observations is done in this case, for difference of mean yields of two varieties.

University	Variety I x_i	Variety II y_i	$d_i = x_i - y_i$	$(d_i - \bar{d})^2$
1	38	45	–7	102.21
2	23	25	–2	26.1121
3	35	31	4	0.7921
4	41	38	3	0.0121
5	44	50	–6	82.9921
6	29	36	–7	102.21
7	37	36	1	4.4521
8	31	40	–9	146.6521
9	38	43	–5	65.7721
				531.2068

$$\bar{d} = \frac{\Sigma d_i}{n} = \frac{-28}{9} = -3.11, \qquad S_d = \sqrt{\frac{\Sigma(d_i - \bar{d})^2}{n-1}} = \sqrt{\frac{531.2068}{9-1}}$$

$S_d = 8.148$

The confidence interval for paired observations of means is

$$\left[\bar{d} - t_{\frac{\alpha}{2}} \frac{S_d}{\sqrt{n}},\ \bar{d} + t_{\frac{\alpha}{2}} \frac{S_d}{\sqrt{n}}\right]$$

$1 - \alpha = 0.95,\ \alpha = 0.05,\ \frac{\alpha}{2} = 0.025\ \gamma = n - 1 = 9 - 1 = 8$

$t_{1-\frac{\alpha}{2}}$ at 8 degrees of freedom = 2.31

$$= \left[-3.11 - (2.31)\frac{8.148}{\sqrt{9}},\ -3.11 + (2.31)\frac{8.148}{3}\right]$$

$$= (-3.11 - 6.273, -3.11 + 6.273)$$

$$= (-9.38, 3.16)$$

EXAMPLE 9.23

A random sample of 20 students obtained a mean of $\bar{x} = 72$ and a variance of $S^2 = 16$ on a college placement test in mathematics. Assuming the scores to be normally distributed, construct a 98% confidence interval for σ^2.

Solution: Given data is

$S^2 = 16$, $\bar{x} = 72$, $n = 20$

The 98% confidence interval for σ^2 is given by

$$\frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2}}} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2}}}$$

where $1 - \alpha = 0.98$,

$\alpha = 0.02$,

$\frac{\alpha}{2} = 0.01$

$\chi^2_{\frac{\alpha}{2}}$ for $n - 1 = 20 - 1 = 19$ degrees of freedom is $\chi^2_{\frac{\alpha}{2}} = 36.2$ and $\chi^2_{1-\frac{\alpha}{2}} = 7.63$

$$= \frac{(20-1)(16)}{36.2} < \sigma^2 < \frac{(20-1)(16)}{7.63}$$

$$= [8.397, 39.84]$$

Hence, 98% confidence interval for σ^2 is (8.397, 39.8).

EXAMPLE 9.24

The following data represents the running times of films produced by two motion picture companies:

Company	Time (in minutes)						
I	103	94	110	87	98	–	–
II	97	82	123	92	175	88	118

(i) Compute 90% confidence interval for the difference between the average running times of films produced by the two companies.

(ii) Construct 90% confidence interval for $\frac{\sigma_1^2}{\sigma_2^2}$. Assume that the running time differences approximately normally distributed with unequal variances.

Solution: From the given data, the mean and variances of the two samples are calculated.

x_i	y_i	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$
103	97	21.16	187.69
94	82	19.36	823.69
110	123	134.56	151.29
87	92	129.96	349.69

(*Continued*)

x_i	y_i	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$
98	175	0.16	4134.49
–	88		515.29
–	118		53.29
492	775	305.2	6215.43

$$\bar{x} = \frac{492}{n_1} = \frac{492}{5} = 98.4$$

$$\bar{y} = \frac{775}{n_2} = \frac{775}{7} = 110.7$$

$$S_1^2 = \frac{305.2}{5-1} = 76.3$$

$$S_2^2 = \frac{6215.43}{7-1} = 1035.9$$

$$\gamma_1, \gamma_2 = n_2 - 1, n_1 - 1$$
$$= (7 - 1, 5 - 1)$$
$$= (6, 4)$$

$$1 - \alpha = 0.90,\ \alpha = 0.01,\ \frac{\alpha}{2} = 0.05,\ 1 - \frac{\alpha}{2} = 0.95$$

$$F_{0.95} = F_{1-\frac{\alpha}{2}}$$

$$F_{0.95}(6, 4) = 6.16$$

$$F_{\alpha}(\gamma_1, \gamma_2) = \frac{1}{F_{1-\alpha}(\gamma_2, \gamma_1)}$$

$$F_{0.05}(6, 4) = \frac{1}{F_{0.95}(4,6)}$$
$$= \frac{1}{4.53}$$
$$= 0.2207$$

Hence, the confidence interval for estimating the ratio of variances is

$$\frac{S_1^2}{s_2^2 F_{\frac{\alpha}{2}}(\gamma_1, \gamma_2)} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{s_1^2}{s_2^2 F_{1-\frac{\alpha}{2}}(\gamma_1, \gamma_2)}$$

$$\frac{1035.9}{76.3(6.16)} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{1035.9}{76.3(0.2207)}$$

$$\frac{1035.9}{470.008} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{1035.9}{16.8394}$$

$$2.204 < \frac{\sigma_1^2}{\sigma_2^2} < 61.516$$

Hence, the confidence interval is (2.2, 61.5).

EXAMPLE 9.25

A random sample of size 26, drawn from a normal population X, has sample variance $S_X^2 = 64$ and a random sample of size 16, drawn from a normal population Y, has a sample variance $S_Y^2 = 100$. Assuming that X and Y are independent, find a 98% confidence interval for $\frac{\sigma_X^2}{\sigma_Y^2}$.

Solution: $n_1 - 1 = 26 - 1 = \gamma_1, \qquad n_2 - 1 = 16 - 1 = \gamma_2$

$\therefore \gamma_1 = 25 \qquad \gamma_2 = 15$

$S_1^2 = 64 \qquad S_2^2 = 100$

$1 - \alpha = 0.98,\ \alpha = 0.02,\ \frac{\alpha}{2} = 0.01,\ 1 - \frac{\alpha}{2} = 0.99$

$$F_{1-\frac{\alpha}{2}}(\gamma_1, \gamma_2) = 3.29, \qquad F_{\frac{\alpha}{2}}(\gamma_1, \gamma_2) = \frac{1}{F_{1-\frac{\alpha}{2}}(\gamma_1, \gamma_2)} = \frac{1}{2.85} = 0.3508$$

The confidence interval for estimating $\frac{\sigma_X^2}{\sigma_Y^2}$ is

$$\Rightarrow \left[\frac{S_1^2}{S_2^2 F_{\frac{\alpha}{2}}(\gamma_1, \gamma_2)}, \frac{S_1^2}{S^2 F_{1-\frac{\alpha}{2}}(\gamma_1, \gamma_2)} \right] = \left[\frac{64}{100(0.3508)}, \frac{64}{100(3.29)} \right]$$

$$= \left(\frac{64}{35.08}, \frac{64}{329} \right) \simeq (1.824, 0.194)$$

Hence, required confidence interval is (0.194, 1.824).

9.12 BAYESIAN ESTIMATION

The methods that were dealt so far in the earlier sections are classical methods of estimation. Now we shall deal with new approach of estimation. The basic difference between classical methods and Bayesian methods is that, in Bayesian concepts, the parameters are viewed as random variables. In this methodology, a distribution is assumed on the parameter θ, which is called 'prior distribution'. This deals with experimenter's prior belief about the parameter θ. All the information about θ from the observed data and prior knowledge are contained in the 'posterior distribution'.

Here are the parameters μ^*, mean and σ^*, the standard deviation of posterior distribution defined, through a theorem.

Theorem

If $\bar{x}$ is the mean of a random sample of size n from a normal population with known variance σ^2, and the prior distribution of the population mean is a normal distribution with mean μ_0 and variance σ_0^2, then the posterior distribution of the population mean is also a normal distribution with mean μ^* and standard deviation σ^* where

$$\mu^* = \frac{n\bar{x}\sigma_0^2 + \mu_0\sigma^2}{n\sigma_0^2 + \sigma^2} \qquad \text{and}$$

$$\sigma^* = \sqrt{\frac{\sigma_0^2\sigma^2}{n\sigma_0^2 + \sigma^2}}$$

Caution:

σ^* is smaller than both σ_0 and $\frac{\sigma}{\sqrt{n}}$ which are the prior standard deviation and the standard deviation of $\bar{x}$, respectively. Hence posterior estimation is more accurate than prior and that of sample data.

Bayesian Interval

In the previous section while computing μ^* and σ^* we have assumed that σ^2 is known. However, this may not be the case always. Hence when σ^2 is not known, we use the sample variance S^2 whenever $n \geq 30$.

Here, the posterior mean μ^* is the Bayesian estimate of the population mean μ.

The $(1 - \alpha)100\%$ Bayesian interval for μ is computed

$$\mu^* - z_{\frac{\alpha}{2}}\,\sigma^* < \mu < \mu^* + z_{\frac{\alpha}{2}}\,\sigma^*$$

This value of μ is centered at the posterior mean μ^* and contains $(1 - \alpha)100\%$ of the posterior probability.

Worked Out Examples

EXAMPLE 9.26

The mean mark in mathematics common entrance test will vary from year to year. If this variation of the mean mark is expressed subjectively by a normal distribution with mean $\mu_0 = 72$ and variance $\sigma_0^2 = 5.76$:

(i) What probability can we assign to the actual mean mark being somewhere between 71.8 and 73.4 for the next year's test?

(ii) Construct a 95% Bayesian interval for μ if the test is conducted for a random sample of 100 students from the next incoming class yielding a mean mark of 70 with standard deviation of 8.

(iii) What posterior probability should we assign to the event of part (i)?

Solution:

(i) Given that $\mu_0 = 72$ and $\sigma_0^2 = 5.76$ the standard deviation $\sigma_0 = 2.4$.

Sample size, $n = 100$

Let $\bar{x}$ be the mean mark in the next year's test.

$$z_1 = \frac{\bar{x} - \mu_0}{\sigma_0}, \text{ when } \bar{x}_1 = 71.8, z_1 = \frac{71.8 - 72}{2.4}$$

$$= -0.083$$

$$\text{when } \bar{x}_2 = 73.4, z_2 = \frac{73.4 - 72}{2.4} = 0.583$$

$$P(\bar{x}_1 < \bar{x} < \bar{x}_2) = P(-0.083 < z < 0.583)$$

$$= P(-0.083 < z < 0) + P(0 < z < 0.583)$$

$$= 0.0319 + 0.2190 = 0.2509$$

(ii) Posterior mean $\mu^* = \dfrac{\eta\bar{x}\sigma_0^2 + \mu_0\sigma^2}{\eta\sigma_0^2 + \sigma^2}$, $\quad \bar{x} = 70$, $\quad \sigma \simeq S = 8$

$$\mu^* = \frac{100(70)(5.76) + 72(8^2)}{100(5.76) + 8^2} = \frac{40320 + 4608}{640}$$

$$= 70.2$$

Posterior standard deviation $\sigma^* = \sqrt{\frac{\sigma^2\sigma_0^2}{n\sigma_0^2 + \sigma^2}}$

$$\sigma^* = \frac{(8^2)(5.76)}{100(5.76) + 8^2}$$

$$= \sqrt{\frac{368.64}{640}}$$

$$= 0.7589$$

$$1 - \alpha = 0.95$$

$$\alpha = 0.05$$

$$\frac{\alpha}{2} = 0.025$$

$$z_{\frac{\alpha}{2}} = 1.96$$

The Bayesian internal for μ is

$$\left[\mu^* - z_{\frac{\alpha}{2}}\sigma^*,\ \mu^* + z_{\frac{\alpha}{2}}\sigma^*\right]$$

$= [70.2 - (1.96)(0.7589),\ 70.2 + (1.96)(0.7589)]$

$= (70.2 - 1.487,\ 70.2 + 1.487) = (68.7,\ 71.68)$

Hence Bayesian interval for μ is (68.7, 71.68)

(iii) The posterior probability for the interval of $\bar{x}(71.8, 73.4)$ when

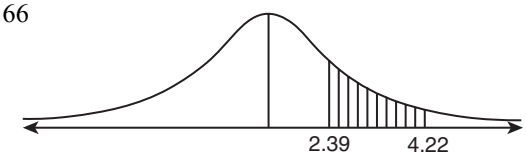
$$\bar{x} = 71.8 \quad z_1 = \frac{\bar{x} - \mu^*}{\sigma^*} = \frac{71.8 - 70.2}{0.7589} = 2.108$$

when $\bar{x} = 73.4,\ z_2 = \frac{\bar{x} - \mu^*}{\sigma^*} = \frac{73.4 - 70.2}{0.7589} = 4.2166$

$$p(71.8 < \bar{x} < 73.4) = p(2.108 < z < 4.22)$$

$$= 0.5 - 0.4826$$

$$= 0.0174$$



Work Book Exercises

1. Many cardiac patients wear implanted pacemakers to control their heartbeat. A plastic connector module mounts on the top of the pacemaker. Assuming a standard deviation of 0.0015 and an approximate normal distribution, find a 95% confidence interval for the mean of an connector modules made by a certain manufacturing company when a random sample of 75 modules has an average of 0.310 inches.
2. A sample of 10 cam shafts intended for use in gasoline engines has an average eccentricity of 1.02 and a standard deviation of 0.044 inches. Assuming that the data may be treated by a random sample from a normal population, determine 95% confidence interval for the actual mean eccentricity of the cam shaft. [**Ans.:** 1.047, 0.993]

3. What is the maximum error one can expect to make with probability 0.90 when using the mean of a random sample of size $n = 64$ to estimate the mean of the population with $\sigma^2 = 2.56$.

[**Ans.:** $E = 0.366$]

4. Assuming that $\sigma = 20.0$, how large a random sample be taken to assert with probability 0.95 that the sample mean will not differ from the true mean by more than 3.0 points?

[**Ans.:** $n = 171$(approximately)]

5. A random sample of 9 metal pieces is taken from a machine that produces them in cylindrical shapes, where diameters are 1.01, 0.97, 1.03, 1.04, 0.99, 0.98, 0.99, 1.01, and 1.03 centimeters. Construct a 99% confidence interval for the mean diameter of pieces from this machine assuming approximate normal distribution.

6. A sample of 150 brand *A* light bulbs showed a mean lifetime of 1,400 hours with a standard deviation of 70 hours. A sample of 200 brand *B* bulbs showed a mean lifetime of 1,200 hours and a standard deviation of 80 hours. Find i) 95% ii) 99% confidence intervals for the mean lifetime of the populations of brands *A* and *B*.

7. If a study showed a sample of $n_1 = 40$ of its bulbs has a mean lifetime of 647 hours of continuous use with a standard deviation of 31 hours, and another study showed a sample of $n_2 = 45$ of its bulbs with a mean life time of 742 hours with standard deviation of 29 hours, construct 95% confidence interval for the difference of their mean lifetimes.

8. In a random sample of 400 claims filed against insurance company writing collision insurance on cars, 112 exceed ₹50,000. Construct a 95% confidence interval for the true proportion of claims filed against this insurance company that exceed ₹50,000.

9. Ten engineering colleges in a country were surveyed. The sample contained 250 electrical engineers, 50 being women; 175 chemical engineers, 40 being women. Construct a 90% confidence interval for the difference between the proportions of women in these two fields of engineering.

10. A random sample of 10 chocolate energy bars of a certain brand has an average of 230 calories with standard deviation of 15 calories.
    (i) Construct 99% confidence interval for the true mean calorie content of this brand of energy bar.
    (ii) Construct 99% confidence interval for σ^2.

11. A taxi company is trying to decide whether to purchase brand *A* or brand *B* tyres for its fleet of taxis. An experiment is conducted using 12 of each brand. The tyres are run until they wear out. The results are as follows:

	$\bar{x}$	S
Brand *A*	36,300 km	5,000 km
Brand *B*	38,100 km	6,100 km

    (i) Compute 95% confidence interval for $\mu_A - \mu_B$.
    (ii) Construct 90% confidence interval for $\frac{\sigma_1^2}{\sigma_2^2}$ assuming approximate normal distribution.

12. The burn time for the first stage of a rocket is a normal random variable with a standard deviation of 0.8 minutes. Assume a normal prior distribution for μ with mean of 8 minutes and a standard deviation of 0.2 minutes. If 10 of these rockets are fired, and the first stage has an average burn time of 9 minutes, find a 95% Bayesian interval for μ.

13. A random sample of 100 teachers in a large metropolitan area revealed a mean weekly salary of ₹487 with a standard deviation of ₹48, with what degree of confidence can we assert that the average weekly salary of all the teachers in the metropolitan area that is between ₹472 and ₹502? [**Ans.:** 99.82%]
14. A random sample of 500 apples was taken from a large consignment and 60 were found to be bad, obtain 98% confidence limits for proportion of bad apples in the consignment. [**Ans.:** 8.5, 15.5]
15. Experience has shown that 10% of a manufactured product is of top quality. What can you say about the maximum error with 95% confidence for 100 items? [**Ans.:** 0.00196]

DEFINITIONS AT A GLANCE

Parameter: A statistical measure of the population such as population mean μ.

Estimation: A procedure by which we can assign values to a population parameter based on the data collected from a sample.

Estimator: A sample statistic which is used to estimate the population parameter.

Point Estimate: A single value of a sample statistic which pertains to the corresponding population parameter.

Confidence Interval: An interval within which the population parameter is likely to fall.

Confidence Level: $(1 - \alpha)100\%$ which gives how much confidence one has so that the true population parameter lies within a confidence interval.

Consistent Estimator: An estimator which gives values more closely approaching the population parameter as the sample size increases.

Unbiased Estimator: An estimator is unbiased when the expected value of the estimator is equal to population parameter that is to be estimated.

Efficient Estimator: An estimator is efficient if the sampling variance of this estimator is less when compared to some other estimator of the population parameter.

Sufficient Estimator: An estimator that uses all the available data pertaining to a parameter.

Degrees of Freedom: The number of values in a sample that can be freely specified once something about a sample is known.

FORMULAE AT A GLANCE

- Point estimate of population mean $\mu = \bar{x}$

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

- Point estimate of population variance, $\sigma^2 = s^2$

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

- Confidence interval for population mean $= \left(\bar{x} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right)$

- Confidence interval for population mean when the population variance is not known $= \left(\bar{x} - t_{\frac{\alpha}{2}} \frac{S}{\sqrt{n}},\ \bar{x} + t_{\frac{\alpha}{2}} \frac{S}{\sqrt{n}} \right)$
- Standard deviation of population proportion is $\sigma_{\hat{P}} = \sqrt{\frac{PQ}{n}}$
- Confidence interval for population proportion is $\left(\hat{p} - z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}\hat{q}}{n}},\quad \hat{p} + z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}\hat{q}}{n}} \right)$
- Confidence interval for difference between two proportion $(P_1 - P_2) - \sqrt{Z_{\frac{\alpha}{2}} \frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}$, $(P_1 - P_2) - \sqrt{Z_{\frac{\alpha}{2}} \frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}$
- Maximum error of estimate in estimating the population mean μ is $e = z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$
- Maximum sample size $n = \left(\frac{z_{\frac{\alpha}{2}} \sigma}{e} \right)^2$
- Maximum error of estimate of proportion P is $e = \sqrt{Z_{\frac{\alpha}{2}} \frac{p(1-p)}{n}}$
- Maximum sample size while estimating the proportion $n = \frac{p(1-p)}{e^2} \left(\frac{z_{\frac{\alpha}{2}}}{e} \right)$
- When the proportion P is unknown, the maximum sample size, $n = \frac{1}{n} \left(\frac{Z_{\frac{\alpha}{2}}}{e} \right)^2$
- Confidence interval for difference of means is $\left[(\bar{x}_1 - \bar{x}_2) - \sqrt{Z_{\frac{\alpha}{2}} \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}, (\bar{x}_1 - \bar{x}_2) + \sqrt{Z_{\frac{\alpha}{2}} \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right]$
- Confidence interval for difference of means when variances are unknown is $\left[(\bar{x}_1 - \bar{x}_2) - t_{\frac{\alpha}{2}} S \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}, (\bar{x}_1 - \bar{x}_2) + t_{\frac{\alpha}{2}} S \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right]$
- Pooled estimate of the unknown common variance is $S^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$
- Confidence interval for difference of means when variances are unequal and unknown is $\left[(\bar{x}_1 - \bar{x}_2) - \sqrt{t_{\frac{\alpha}{2}} \frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}, (\bar{x}_1 - \bar{x}_2) + \sqrt{t_{\frac{\alpha}{2}} \frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \right]$
- Confidence interval for difference between means for paired observations is $\left[\bar{d} - t_{\frac{\alpha}{2}} \frac{S_d}{\sqrt{n}}, \bar{d} + t_{\frac{\alpha}{2}} \frac{S_d}{\sqrt{n}} \right]$

- When the observations are paired, mean of the differences, $\bar{d} = \sum \frac{d_i}{n}$ where $d_i = x_i - y_i$
- Standard deviation of the differences $S_d = \sqrt{\frac{\sum (d_i - \bar{d})^2}{n-1}}$
- Confidence interval for variance is $\left[\frac{(n-1)S^2}{\lambda^2_{\frac{\alpha}{2}}}, \frac{(n-1)S^2}{\lambda^2_{1-\frac{\alpha}{2}}} \right]$
- Confidence interval for ratio of two variances is $\left[\frac{S_1^2}{S_2^2 F_{\frac{\alpha}{2}}(\gamma_1, \gamma_2)}, \frac{S^2}{S_2^2 F_{1-\frac{\alpha}{2}}(\gamma_1, \gamma_2)} \right]$
- In Bayesian estimation, the posterior:
  - Population mean $\mu* = \frac{n\bar{x}\sigma_0^2 + \mu_0\sigma^2}{n\sigma_0^2 + \sigma^2}$
  - Posteriors deviation $\sigma* = \sqrt{\frac{\sigma_0^2\sigma^2}{n\sigma_0^2 + \sigma^2}}$
- Bayesian interval is $(\mu* - Z_{\frac{\alpha}{2}}\sigma*, \mu* + Z_{\frac{\alpha}{2}}\sigma*)$

OBJECTIVE TYPE QUESTIONS

1. If we can assert with 95% confidence that the maximum error is 0.05 and $P = 0.2$ then, the sample size is _____.
   (a) 242 (b) 240
   (c) 246 (d) 250
2. A random sample of size 100 has a standard deviation of 5. Then the maximum error with 95% confidence is _____.
   (a) 0.98 (b) 0.098
   (c) 9.8 (d) 0.0098
3. A random sample of size 81 was taken whose variance is 20.25 and mean is 32, then the confidence interval for mean is _____.
   (a) (30.83, 30.16) (b) (30.83, 33.16)
   (c) (33.16, 34.16)) (d) (30.83, 40.83)
4. If $n = 144$, $\sigma = 4$ and $\bar{x} = 150$, then 95% confidence interval for mean is _____.
   (a) (149.35, 149.65) (b) (143.35, 149.65)
   (c) (143.35, 143.65) (d) (149.35, 150.65)
5. In a sample of 500 people in Maharashtra, 300 are wheat eaters. Then the maximum error with 99% confidence is _____.

(a) 0.56 (b) 0.056
(c) 0.0056 (d) 5.6

6. In Bayesian estimation, the posterior standard deviation is _____.

(a) $\sigma^* = \sqrt{\frac{\sigma_0^2 \sigma^2}{n\sigma_0^2 + \sigma^2}}$ (b) $\sigma^* = \sqrt{\frac{\sigma^2}{n\sigma_0^2 + \sigma^2}}$

(c) $\sigma^* = \sqrt{\frac{\sigma_0^2}{\sigma_0^2 + n\sigma^2}}$ (d) $\sigma^* = \sqrt{\frac{\sigma_0^2 \sigma^2}{\sigma_0^2 + n\sigma^2}}$

7. Confidence interval for variance is given by _____.

(a) $\left[\frac{ns^2}{\lambda^2_{\frac{\alpha}{2}}}, \frac{ns^2}{\lambda^2_{\left(1-\frac{\alpha}{2}\right)}}\right]$ (b) $\left[\frac{(n-1)s^2}{\lambda^2_{\frac{\alpha}{2}}}, \frac{(n-1)s^2}{\lambda^2_{1-\frac{\alpha}{2}}}\right]$

(c) $\left[\frac{s^2}{n\lambda_{\frac{\alpha}{2}}}, \frac{s^2}{n\lambda_{\left(1-\frac{\alpha}{2}\right)}}\right]$ (d) $\left[\frac{s^2}{(n-1)\lambda^2_{\frac{\alpha}{2}}}, \frac{s^2}{(n-1)\lambda^2_{1-\frac{\alpha}{2}}}\right]$

8. Maximum error of estimate in estimating the population mean μ is _____.

(a) $e = z_{\frac{\alpha}{2}} \frac{\sigma}{n}$ (b) $e = z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$

(c) $e = \frac{z_{\frac{\alpha}{2}}}{\frac{\sigma}{\sqrt{n}}}$ (d) $z_\alpha \frac{\sigma}{\sqrt{n}}$

9. The number of values in a sample that can be freely specified once something about a sample is known is _____.

(a) Sample size (b) Parameter
(c) Degrees of freedom (c) Statistic

10. An estimator which gives more closely approaching the population parameter as the sample size increases is called _____.

(a) Efficient estimator (b) Sufficient estimator
(c) Unbiased estimator (d) Consistent estimator

ANSWERS

1. (c) 2. (a) 3. (b) 4. (d) 5. (b) 6. (a) 7. (b) 8. (b)
9. (c) 10. (d)

10 Curve Fitting

Prerequisites

Before you start reading this unit, you should:

- Have some knowledge in partial differentiation
- Be able to do some calculations like summations, logarithms of some numbers and their anti logarithms.
- Have little knowledge in plotting of points

Learning Objectives

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

- Know fitting a curve to a data
- Obtain normal equations for any curve
- Understanding the method of least squares in fitting any curve

INTRODUCTION

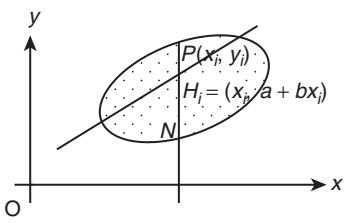
The general problem in curve fitting is to find an analytical expression to a given set of data points (x_i, y_i), where x is the independent variable and y is the dependent variable.

In practical statistics, the curve fitting enables us to represent the relationship between two variables such as polynomials, logarithmic functions, and exponential functions. Theoretically this is useful for the study of correlation and regression.

Suppose a set of data points (x_i, y_i), $i = 1, 2 \ldots n$ are given where they satisfy the equations $y_i = f(x_i)$. When these points are plotted the diagram so obtained is called a scattered diagram or scatter plot.

10.1 THE METHOD OF LEAST SQUARES

This is also called as Legendre's principle of least squares. This consists of minimizing the sum of the squares of the deviations of the actual values of y from their estimated values which is given by the line of best fit. The procedure of finding the equation of a line which best fits a given set of paired data, is called the method of least squares.



Line of best fit

10.2 FITTING OF A STRAIGHT LINE

Consider the straight line,

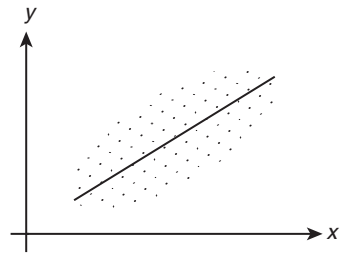
$$y = a + bx \tag{10.1}$$

to a set of n points (x_i, y_i), $i = 1, 2, 3, \ldots n$. The above equation represents a family of straight lines for different values of the parameters a and b.

The problem here is to find a and b so that the line so obtained is the line of best fit.

Let $P_i(x_i, y_i)$ be any general point in the scattered diagram. Draw a perpendicular from P_i to the x-axis $P_i N$, so that it intersects the line at the point H_i. Since H_i lies on the straight line its x-coordinate is x_i and y-coordinate is $a + bx_i$ Hence the point $H_i = (x_i, a + bx_i)$

We get many straight lines for the set of given data points. From these lines, we pick the one that is taking maximum number of points on it or best fits to the data points. To determine how close the i^{th} data point is to the estimated line, we should measure the vertical distance from the data point to this line. This distance is called i^{th} residual or error of estimate.



Scattered diagram

$$P_i H_i = P_i N - H_i N$$
$$= y_i - (a + bx_i)$$

The error of estimate is $e_i = \sum \{y_i - (a + bx_i)\}^2$.

From the principle of least squares and theory of calculus, the partial derivatives of e_i with respect to the parameters a and b vanish.

Hence,

$$\frac{\partial e_i}{\partial a} = 0 \Rightarrow -2\sum \{y_i - (a + bx_i)\} = 0$$

$$\frac{\partial e_i}{\partial b} = 0 \Rightarrow -2\sum \{y_i - (a + bx_i)(x_i)\} = 0$$

$$\sum y_i - \{n\ a + b\sum x_i\}$$

$$\sum y_i x_i - \{a\sum x_i + b\sum x_i^2\}$$

Hence the normal equations for estimating the parameters are

$$\sum y_i = n\ a + b\sum x_i \tag{10.2}$$

$$\sum y_i x_i = a\sum x_i + b\sum x_i^2 \tag{10.3}$$

The quantities $\sum y_i$, $\sum x_i$, $\sum x_i^2$, and $\sum y_i x_i$ can be obtained from the given data points (x_i, y_i). Then the above three equations involve three unknowns a, b, and c which can be obtained on solving them.

With the values of a, b, and c so obtained gives the required second degree parabola of best fit to the given set of data points is $y = a + bx$.

Worked Out Examples

EXAMPLE 10.1

The following data pertains to the number of jobs per day and the central processing unit (CPU) time required. Fit a straight line. Estimate the mean CPU time at $x = 3.5$.

No. of jobs	1	2	3	4	5
CPU time	2	5	4	9	10

(JNTU 2001, 2004s, 2005, 2005s, 2007)

Solution: Let the straight line which is to be fit to the given data is $y = a + bx$. To estimate the parameters a and b the normal equations are

$$\sum_{i=1}^{n} y_i = n\ a + b\sum_{i=1}^{n} x_i \tag{10.4}$$

$$\sum_{i=1}^{n} x_i y_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 \tag{10.5}$$

x_i	y_i	$x_i y_i$	x_i^2
1	2	2	1
2	5	10	4
3	4	12	9
4	9	36	16
5	10	50	25
$\sum x_i = 15$	$\sum y_i = 30$	$\sum x_i y_i = 110$	$\sum x_i^2 = 55$

Substituting the values from the table, in equations (10.1) and (10.2) we get

$$30 = 5a + 15b \tag{10.6}$$

$$110 = 15a + 55b \tag{10.7}$$

$$6 = a + 3b \quad \{\text{by dividing (10.6) by 5}\} \tag{10.8}$$

$$22 = 3a + 11b \quad \{\text{by dividing (10.7) by 5}\} \tag{10.9}$$

$18 = 3a + 9b$ {multiplying (10.6) with 3 and subtracting from (10.9)}, we get

$4 = 2b$

$\because\ b = 2$

$a = 6 - 3b$

$= 6 - 3(2) = 0$

Hence, the required straight line is $y = 2x$

Hence the estimate of mean CPU time for $x = 3.5$ is $y = 2\ (3.5) = 7$.

EXAMPLE 10.2

The following are measurements of the air, velocity, and evaporation coefficient of burning fuel droplets in air impulse engine. Fit a straight line to the above data by the method of least squares.

Air velocity, x	20	60	100	140	180	220	260	300	340	380
Evaporation coefficient, y	0.18	0.37	0.35	0.78	0.56	0.75	1.18	1.36	1.17	1.65

Solution: Let the straight line which is to be fit to the given data is $y = a + bx$. To estimate the parameters a and b the normal equations are

$$\sum_{i=1}^{n} y_i = n\,a + b\sum_{i=1}^{n} x_i \tag{10.10}$$

$$\sum_{i=1}^{n} x_i y_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 \tag{10.11}$$

x_i	y_i	$x_i y_i$	x_i^2
20	0.18	3.6	400
60	0.37	22.2	3600
100	0.35	35.0	10000
140	0.78	109.2	19600
180	0.56	100.8	32400
220	0.75	165.0	48400
260	1.18	306.8	67600
300	1.36	408.0	90000
340	1.17	397.8	115600
380	1.65	627.0	144400
$\sum x_i = 2000$	$\sum y_i = 8.35$	$\sum x_i y_i = 2175.4$	$\sum x_i^2 = 532000$

Substituting the above values from the table, in equations (10.10) and (10.11) we get

$$8.35 = 10a + 2000b \tag{10.12}$$

$$2175.4 = 2000a + 53200b \tag{10.13}$$

Multiplying (10.12) by 200 and subtracting from equation (10.13), we get

$$1670 = 2000a + 400000b$$

$$2175.4 = 2000a + 53200b, \text{ we get}$$

$$505.4 = 132000b$$

$$b = \frac{505.4}{132000}$$

$$= 0.00383$$

Substituting b in (10.12) we get

$$10a = 8.35 - 2000b$$

$$= 8.35 - 2000\,(0.00383)$$

$$= 0.069$$

$\therefore$ The line of best fit is $y = 0.069 + 0.00383x$

10.3 FITTING OF A SECOND DEGREE PARABOLA

Let (x_i, y_i) be the given set of data points. Let $y = a + bx + cx^2$ be the second degree parabola which is to be fit to the given data points.

The error of estimate is given by

$$e_i = \sum \{y_i - (a + bx_i + cx_i^2)\}^2$$

To estimate the parameters a, b, and c, from the theory of calculus, the partial derivatives with respect to the a, b, and c should vanish.

$$\frac{\partial e_i}{\partial a} = 0 \Rightarrow -2\sum \{y_i - (a + bx_i + cx_i^2)\} = 0$$
$$\frac{\partial e_i}{\partial a} = 0 \Rightarrow -2\sum \{y_i - (a + bx_i + cx_i^2)\}(x_i) = 0$$
$$\frac{\partial e_i}{\partial a} = 0 \Rightarrow -2\sum \{y_i - (a + bx_i + cx_i^2)\}(x_i^2) = 0$$

Hence, the normal equations for estimating the parameters are

$$\sum y_i = n\ a + b\sum x_i + c\sum x_i^2 \tag{10.14}$$

$$\sum y_i x_i = a\sum x_i + b\sum x_i^2 + c\sum x_i^3 \tag{10.15}$$

$$\sum y_i x_i^2 = a\sum x_i^2 + b\sum x_i^3 + c\sum x_i^4 \tag{10.16}$$

The quantities, $\sum y_i$, $\sum x_i$, $\sum x_i^2$, $\sum x_i^3$, $\sum x_i^4$, $\sum y_i x_i$, and $\sum y_i x_i^2$ can be obtained from the given data points (x_i, y_i). Then the above three equations involve three unknowns a, b, and c which can be obtained on solving them.

We get the values of a, b, and c so that $y = a + bx + cx^2$ is the best fitting parabola of second degree to the given set of data points.

Worked Out Examples

EXAMPLE 10.3

Fit a second degree parabola to the following data using the method of least squares.

x	0	1	2	3	4
y	1	1.8	1.3	2.5	6.3

(JNTU 1999, 2005, 2007)

Solution: Let the second degree parabola which is to be fit to the given data is $y = a + bx + cx^2$. To estimate the parameters a, b, and c, the normal equations are

$$\sum_{i=1}^{n} y_i = na + b\sum_{i=1}^{n} x_i + c\sum_{i=1}^{n} x_i^2 \tag{10.14}$$

$$\sum_{i=1}^{n} x_i y_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 + c\sum_{i=1}^{n} x_i^3 \tag{10.15}$$

$$\sum y_i x_i^2 = a\sum x_i^2 + b\sum x_i^3 + c\sum x_i^4 \tag{10.16}$$

x_i	y_i	$x_i y_i$	x_i^2	$y_i x_i^2$	x_i^3	x_i^4
0	1.0	0	0	0	0	0
1	1.8	1.8	1	1.8	1	1
2	1.3	2.6	4	5.2	8	16
3	2.5	7.5	9	22.5	27	81
4	6.3	25.2	16	100.8	64	256
$\sum x_i = 10$	$\sum y_i = 12.9$	$\sum x_i y_i = 37.1$	$\sum x_i^2 = 30$	$\sum y_i x_i^2 = 130.3$	$\sum x_i^3 = 100$	$\sum x_i^4 = 354$

Substituting these values in the above normal equations we get

$$12.9 = 5a + 10b + 30c \tag{10.17}$$

$$37.1 = 10a + 30b + 100c \tag{10.18}$$

$$130.3 = 30a + 100b + 354c \tag{10.19}$$

Multiplying (10.17) by 2 and subtracting from (10.18) we get,

$$\begin{aligned} 25.8 &= 10a + 20b + 60c \\ 37.1 &= 10a + 30b + 100c \\ \hline 11.3 &= 10b + 40c \end{aligned} \tag{10.20}$$

$$111.3 = 30a + 90b + 300c \tag{10.21}$$

$$130.3 = 30a + 100b + 354c \tag{10.19}$$

$$19 = 10b + 54c \tag{10.22}$$

$$11.3 = 10b + 40c \tag{10.20}$$

$$7.7 = 14c$$

$$c = \frac{7.7}{14}$$

$$= 0.55$$

Substituting the value of c in equation (10.20), we get

$$11.3 = 10b + 40(0.55)$$

$$b = \frac{11.3 - 40(0.55)}{10}$$

$$= \frac{11.3 - 22}{10}$$

$$b = -1.07$$

Substituting the values of b and c in equation (10.17), we get

$$12.9 = 5a + 10(-1.07)b + 30\,(0.55)$$

$$a = \frac{12.9 + 10(1.07) - 30(0.55)}{5}$$

$$= \frac{12.9 + 10.7 - 16.5}{5}$$

$$a = 1.42$$

Hence, the parabola that is obtained with the values of a, b, and c is $y = 1.42 - 1.07x + 0.55x^2$.

10.4 FITTING OF EXPONENTIAL CURVE AND POWER CURVE

Fitting of Power Curve: $Y = a\,X^b$

Let the power curve that is to be fit to a set of data points (x_i, y_i), $i = 1, 2, \ldots n$ be

$$Y = a\,X^b \tag{10.23}$$

Taking logarithm on both sides of (10.23), we get

$$\log Y = \log a + b \log X$$

$$U = A + bV \tag{10.24}$$

where $U = \log Y$, $A = \log a$, $V = \log X$

Equation (10.24) is a linear equation in V and U.

Hence, the normal equations for estimating the parameters A and b are

$$\sum U = n\,A + b\sum V \tag{10.25}$$

$$\sum UV = A\sum V + b\sum V^2 \tag{10.26}$$

Equations (10.25) and (10.26) are normal equations for solving A and b.

However, since $A = \log a$, $a =$ antilog (A) with the values of a and b so obtained equation (10.23) gives the curve of best fit to the given set of data points.

Fitting of Exponential Curve: $Y = a\,b^X$

Let the exponential curve that is to be fit to a set of data points (x_i, y_i), $i = 1, 2, \ldots n$ be

$$Y = a\,b^X \tag{10.27}$$

Taking logarithm on both sides of (10.27), we get

$$\log Y = \log a + X \log b$$

$$U = A + BX \tag{10.28}$$

where $U = \log Y$, $A = \log a$, $B = \log b$.

This is a linear equation in U and X. The normal equations for estimating the parameters A and B are

$$\sum U = nA + B\sum X \tag{10.29}$$

$$\sum UX = A\sum X + B\sum X^2 \tag{10.30}$$

Equations (10.29) and (10.30) are normal equations for solving A and B.

However, since $A = \log a$ and $B = \log b$

a = antilog (A), b = antilog (B)

With the values of a and b, so obtained equation (10.27) gives the curve of best fit to the given set of data points.

Fitting of Exponential Curve: $Y = a\, e^{bX}$

Let the exponential curve that is to be fit to a set of data points (x_i, y_i), $i = 1, 2, \ldots. n$ be

$$Y = a\, e^{bX} \tag{10.31}$$

Taking logarithm on both sides of (10.31), we get

$$\log Y = \log a + bX \log e$$

$$= \log a + X\,(b \log e)$$

$$U = A + BX \tag{10.32}$$

where $U = \log Y$, $A = \log a$, $B = b \log e$.

This is a linear equation in U and X. The normal equations for estimating the parameters A and B are

$$\sum U = nA + B\sum X \tag{10.33}$$

$$\sum UX = A\sum X + B\sum X^2 \tag{10.34}$$

Equations (10.33) and (10.34) are normal equations for solving A and B.

However, since $A = \log a$ and $B = b \log e$

a = antilog (A), $b = \dfrac{B}{\log e}$

With the values of a and b, so obtained equation (10.31) gives the curve of best fit to the given set of data points.

Worked Out Examples

EXAMPLE 10.4

Fit an exponential curve of the form $Y = a\, b^X$ to the following data:

x	1	2	3	4	5	6	7	8
y	1.0	1.2	1.8	2.5	3.6	4.7	6.6	9.1

Solution: Let $Y = a\,b^X$ be the curve which is to be fit to the given set of points.

Taking logarithm on both sides of above equation we get

$$\log Y = \log a + X \log b$$

$$U = A + BX \tag{10.35}$$

where $U = \log Y$, $A = \log a$, $B = \log b$

X	Y	$U = \log Y$	XU	X^2
1	1.0	0.0000	0.0000	1
2	1.2	0.0792	0.1584	4
3	1.8	0.2553	0.7659	9
4	2.5	0.3979	1.5916	16
5	3.6	0.5563	2.7815	25
6	4.7	0.6721	4.0326	36
7	6.6	0.8195	5.7365	49
8	9.1	0.9590	7.6720	64
$\sum X = 36$	$\sum Y = 30.5$	$\sum U = 3.7393$	$\sum XU = 22.7385$	$\sum X^2 = 204$

The normal equations for estimating the parameters A and B are

$$\sum U = nA + B\sum X \tag{10.33}$$

$$\sum UX = A\sum X + B\sum X^2 \tag{10.34}$$

Substituting the values from the table in the equations (10.33) and (10.34) we get,

$$3.7393 = 8A + 36B \tag{10.36}$$

$$22.7385 = 36A + 204B \tag{10.37}$$

Solving (10.36) and (10.37) we get,
$B = 0.1408$ and $A = -0.1662$
$B = \text{antilog}\,(B) = 1.383$ and $a = \text{antilog}\,(A) = 0.6821$
Hence, the required equation is $Y = 0.6821\,(1.383)^x$

EXAMPLE 10.5

For the data given below, find the equation to the best fitting exponential curve of the form $Y = a\,e^{bX}$

X	1	2	3	4	5	6
Y	1.6	4.5	13.8	40.2	125.0	300.0

Solution: Let the exponential curve that is to be fit to a set of data points (x_i, y_i), $i = 1, 2, \ldots n$ be

$$Y = a\, e^{bX} \tag{10.38}$$

Taking logarithm on both sides of (10.38), we get

$$\log Y = \log a + bX \log e$$

$$= \log a + X\,(b \log e)$$

$$U = A + BX \tag{10.39}$$

where $U = \log Y$, $A = \log a$, $B = b \log$ e

X	Y	$U = \log Y$	XU	X^2
1	1.6	0.20412	0.20412	1
2	4.5	0.653213	1.306425	4
3	13.8	1.139879	3.419637	9
4	40.2	1.604226	6.416904	16
5	125.0	2.09691	10.48455	25
6	300.0	2.477121	14.86273	36
$\sum X = 21$	$\sum Y = 485.1$	$\sum U = 8.175469$	$\sum XU = 36.69436$	$\sum X^2 = 91$

The normal equations for solving (10.39) are

$$\sum U = n\,A + B \sum X \tag{10.40}$$

$$\sum UX = A \sum X + B \sum X^2 \tag{10.41}$$

$$8.175469 = 6A + B(21) \tag{10.42}$$

$$36.69436 = A(21) + B(91) \tag{10.43}$$

Multiplying equation (10.42) by 21 and (10.43) by 6 and subtracting, we get

$$171.6848 = 126A + 441B$$

$$220.1662 = 126A + 546B$$

$$-48.4813 = -105B$$

$$B = \frac{48.4813}{105}$$

$$= 0.461727$$

Substituting the value of B in equation (10.42) we get,

$$8.175469 = 6A + (0.461727)\ (21)$$

$$A = \frac{8.175469 - (0.461727)(21)}{6}$$

$$= -0.25347$$

$$a = \text{antilog}\,(-0.25347) = 0.5578708$$

$$B = b \log e$$

$$b = \frac{B}{\log e} = \frac{0.461727}{\log e}$$

$$= \frac{0.461727}{0.43429}$$

$$= 1.0631767$$

$Y = (0.5578708)\, e^{(1.0631767)X}$ is the curve of best fit to the given data.

EXAMPLE 10.6

Derive the least square equations for fitting a curve of the type $Y = aX + \left(\frac{b}{X}\right)$ to a set of data points (x_i, y_i), $i = 1, 2, 3 \ldots n$

Solution: To fit a curve of the given type is to estimate the parameters a and b from the theory of principle of least squares.

The error of estimate for the above curve is

$$e_i = \sum_{i=1}^{n} \left\{ y_i - ax_i - \frac{b}{x_i} \right\}^2$$

According to the principle of least squares and theory of calculus, the partial derivatives with respect to the parameters a and b should vanish.

$$\frac{\partial e_i}{\partial a} = 0 \Rightarrow -2\sum_{i=1}^{n} \left\{ y_i - ax_i - \frac{b}{x_i} \right\} (x_i$$

$$\frac{\partial e_i}{\partial b} = 0 \Rightarrow -2\sum_{i=1}^{n} \left\{ y_i - ax_i - \frac{b}{x_i} \right\} \left(\frac{1}{x_i} \right)$$

On simplifying the above equations we get,

$$\sum_{i=1}^{n} x_i y_i = a\sum_{i=1}^{n} x_i^2 + n\, b \tag{10.44}$$

$$\sum_{i=1}^{n} \frac{y_i}{x_i} = n\, a + b\sum_{i=1}^{n} \frac{1}{x_i^2} \tag{10.45}$$

These are called normal equations for estimating the parameters a and b. The quantities $\sum_{i=1}^{n} x_i y_i$, $\sum_{i=1}^{n} x_i^2$, $\sum_{i=1}^{n} \frac{y_i}{x_i}$, and $\sum_{i=1}^{n} \frac{1}{x_i^2}$ can be obtained from the given data points.

Substituting them in the normal equations (10.44) and (10.45) and solving for a and b we get the equation of best fit as $Y = aX + \left(\frac{b}{X}\right)$.

Work Book Exercises

1. Fit a straight line to the following data:

X	1	2	3	4	5	6	7	8	9
Y	9	8	10	12	11	13	14	16	15

2. Fit a second degree parabola to the following data:

X	10	12	15	23	20
Y	14	17	23	25	21

3. Fit a curve of the form $Y = a\,X^b$ to the following data by the method of least squares:

X	1	2	3	4	5
Y	0.5	2	4.5	8	12.5

4. Fit the curve $Y = a\,e^{bX}$ to the following data:

X	77	100	185	239	285
Y	2.4	3.4	7.0	11.1	19.6

5. The following table shows how many weeks a sample of six persons have worked at an automobile inspection station and the number of cars each one inspected between noon and 2 pm on a given day:

X	2	7	9	1	5	12
Y	13	21	23	14	15	21

Fit a straight line by least square method. Estimate how many cars someone who has been working at the inspection station for 8 weeks can be expected to inspect during the year in 2 hours period.

[**Ans.:** $y = 15.85 + 0.33x$, $y = 18.49$]

6. Fit a second degree parabola $y = a + bx + cx^2$ to the following data:

X	1	2	3	4	5	6	7	8	9
Y	2	6	7	8	10	11	11	10	9

[**Ans.:** $y = -1 + 3.55x - 0.27x^2$]

7. Fit a curve of best fit of the type $Y = a\,e^{bX}$ to the following data by the method of least squares: (JNTU 2000)

X	1	5	7	9	12
Y	10	15	12	15	21

[**Ans.:** $y = 9.4754\ e^{0.059x}$]

8. Fit a curve of the type $Y = a\ X^b$ to the following data:

X	1	2	3	4	5	6
Y	2.98	4.26	5.21	6.10	6.80	7.50

[**Ans.:** $y = 2.978,\ x^{0.5142}$]

DEFINITIONS AT A GLANCE

Method of Least Squares: This consists of minimizing the sum of the squares of the deviations of the actual values of y from their estimated values.

Scattered Diagram: Suppose a set of data points (x_i, y_i), $i = 1, 2 \ldots n$ are given where they satisfy the equation $y_i = f(x_i)$, when these points are plotted the diagram so obtained is called a scattered diagram or scatter plot.

Error of Estimate: To determine how close the i^{th} data point is to the estimated line, we measure the vertical distance from the data point to this line. This distance is called i^{th} residual or error of estimate.

FORMULAE AT A GLANCE

- The normal equations for estimating the parameters a and b for fitting a straight line:

$$y = a + bx \text{ are } \sum y_i = na + b\sum x_i,\ \sum y_i x_i = a\sum xi + b\sum x_i^2$$

- The normal equations for estimating the parameters a, b, and c for fitting a second parabola:

$$y = a + bx + cx^2 \text{ are } \sum y_i = na + b\sum x_i + c\sum x_i^2,\ \sum y_i x_i = a\sum x_i + b\sum x_i^2 + c\sum x_i^3$$

$$\sum y_i x_i^2 = a\sum x_i^2 + b\sum x_i^3 + c\sum x_i^4$$

- To fit a power curve: $Y = a\ X^b$ to a set of data points (x_i, y_i), $i = 1, 2, \ldots n$,

$$U = A + b\ V, \text{ where } U = \log Y,\ A = \log a,\ V = \log X$$

Hence, the normal equations for estimating the parameters A and b are

$$\sum U = n\,A + b\sum V$$

$$\sum UV = A\sum V + bV^2$$

OBJECTIVE TYPE QUESTIONS

1. The normal equations for fitting a straight line $y = a + bx$ to a set of data points (x_i, y_i) are _____.

   (a) $\sum y_i = na + b\sum x_i$
   $\sum y_i x_i = a\sum x_i + b\sum x_i^2$

   (b) $ny = a + b\sum x_i$
   $\sum y_i x_i = na + b\sum x_i^2$

   (c) $\sum y_i = a + nb$
   $\sum y_i x_i = na + b\sum x_i^2$

   (d) none

2. The normal equations for fitting a curve of the type $Y = a\,b^X$ are _____ to a set of data points (x_i, y_i).

   (a) $\sum U = \sum X + nB$
   $\sum UX = nA + B\sum X$

   (b) $\sum U = nA + B\sum X$
   $\sum UX = A\sum X + B\sum X^2$

   (c) $nU = A + B\sum X$
   $\sum UX = nA + B\sum X^2$

   (d) none, where $U = \log y$
   $A = \log A$
   $B = \log b$

3. The normal equations for fitting a curve of the type $Y = a\,e^{bX}$ are _____ to a set of data points (x_i, y_i).

   (a) as in 2

   (b) as in 2

   (c) as in 2

   (c) as in 2, where $U = \log y$
   $A = \log A$
   $B = b\log e$

4. The normal equations for fitting a curve of the type $Y = a\,X^b$ are _____ to a set of data points (x_i, y_i).

   (a) $\sum U = nA + b\sum V$
   $\sum UV = A\sum V + b\sum V^2$

   (b) $\sum U = nA + b\sum V$
   $\sum UV = nA + b\sum V^2$

   (c) $\sum U = \sum A + nb$
   $\sum UV = nA + b\sum V^2$

   (d) none

5. The method of minimizing the sum of squares of the deviations from the observations is called _____.

   (a) Optimization method
   (b) Graphical method
   (c) Method of least squares
   (d) None

6. The diagram in which the data points are plotted or scattered around a curve is called _____.

   (a) Plotted diagram
   (b) Scattered diagram
   (c) Graphical diagram
   (d) None

ANSWERS

1. (a) 2. (b) 3. (b) 4. (a) 5. (c) 6. (b)

11 Correlation

Prerequisites

Before you start reading this unit, you should:

- Know calculating mean and standard deviation for a given data
- Know some basic ideas of plotting graph and draw inference from a graph

Learning Objectives

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

- Understand the importance and limitations of correlation analysis
- Know the difference between linear and nonlinear correlation, positive and negative and simple, partial, and multiple correlation
- Know how and whether variables are correlated through shatter display
- Calculate correlation coefficient for univariate and bivariate distributions

INTRODUCTION

In the earlier chapters, we have seen variables that are dependent on one another. Now in this chapter let us look at the extent to which two variables are related. Sometimes we can examine the relationships between price and demand, weight and strengths, input and output of waste water treatment plant, the tensile strength and the hardness of aluminium, etc.

Some Definitions

Some definitions of the term 'correlation' as given by Economists and Statisticians are as follows:

- Correlation means that between two series or groups of data there exists some casual connection.

  —W. I. King
- Correlation analysis attempts to determine the degree of relationship between the variables.

  —Ya Lun Chou
- When a relationship is of a quantitative nature, the appropriate statistical tool for discovering and measuring the relationship and expressing it in a brief formula is known as correlation.

  —Croxton and Cowden

As a summary to the above definitions, the correlation can be defined as the statistical tool or technique which measures and analyses the degree or extent to which two or more variables vary with reference to one another.

To know if there is any correlation between two variables under study, we need to know if the change in one variable causes a change in the other variable.

11.1 TYPES OF CORRELATION

Now, let us look at the different types of correlation. The four types of correlation are as follows:

1. **Positive and negative correlation:** If the two variables under study deviate in the same direction, that is, if the increase (or decrease) in one variable affects on an increase (or decrease) in the other variable, the two variables are said to be positively or directly correlated. As an example correlation between heights and weights of a group of persons is positive.

   If the two variables under study deviate in opposite directions, that is, if an increase (or decrease) in one variable affects on a decrease (or increase) in the other variable, the two variables are said to be negatively or diversely correlated.

   As an example, the correlation between the volume and pressure of a perfect gas is negative.

2. **Simple and multiple correlation:** Simple correlation corresponds to the correlation between only two variables. The correlation between yield of rice or wheat and use of chemical fertilizers is simple.

   Multiple correlation corresponds to correlation between more than two variables. Considering the same example as above, the correlation between yield of wheat, use of fertilizer, and use of a pesticide is multiple as these variables are under study.

3. **Partial and total:** The correlation between two variables excluding other variables included for calculation of total correlation is partial correlation. For example, if we are looking at the correlation between yield of wheat and the use of chemical fertilizers, excluding the effect of pesticides and manures then it is called the partial correlation.

   The total correlation is based on all the relevant variables, which is normally not feasible.

4. **Linear and non-linear:** The correlation between two variables is linear when variations in the values of the two variables have a constant ratio. Generally, correlation refers to linear correlation and the graph of such linear correlation is a straight line.

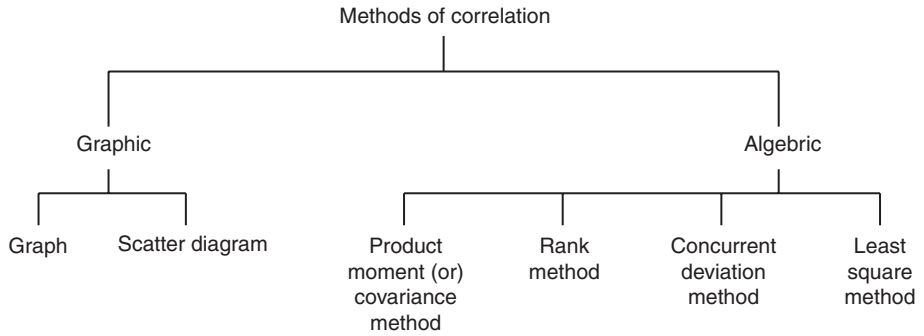
   The non-linear correlation refers to curvilinear correlation where the amount of change in one variable does not bear a constant ratio to the amount of change in the other variable.

   For example, consider the previous example, if we double the use of fertilizers, the yield of wheat need not be doubled (yield may increase, but not in the same proportion).

After knowing the different types of correlation, now let us look at various methods of studying correlation.

11.2 METHODS OF CORRELATION

The different methods of correlation are represented in the form of a chart, given as follows:

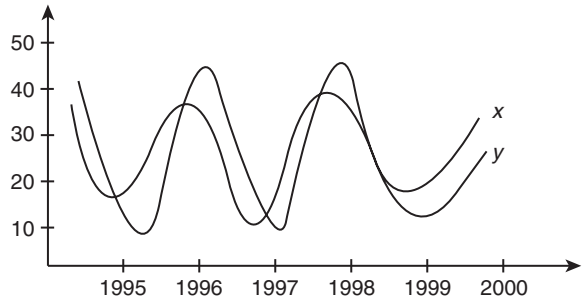


Different methods of correlation

Graphic Method of Correlation

1. **Graphs:** The data related to two series X and Y are plotted on a graph. By observing the direction and closeness of the two curves, we can know the extent of correlation among the two curves. If the two curves move in the same direction, the correlation is positive. If they move in different directions, the correlation between them is negative. If the two curves show no definite pattern then there is very low degree of correlation between the variables.

   Let us consider the following example:



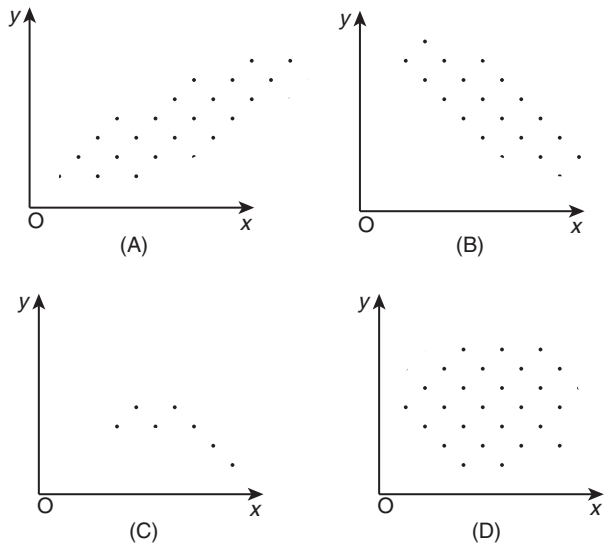
An example of graphic correlation

   The above example shows that X and Y are closely related and hence the correlation among them is positive.

2. **Scatter diagram:** This is a diagrammatic representation of the correlation between the two variables for a bivariate data. If (x_i, y_i) represents the bivariate data, let the value x_i be represented along x-axis and the value y_i be represented along y-axis. The diagram of dots or points (x_i, y_i) is called scatter diagram.

   This gives us a clear idea of whether the variables under study are correlated or not.

   If the points are very dense, that is, close to each other, then there is a good amount of correlation between the two variables. On the other hand, if the points are widely scattered, then there is a poor correlation between the variables.



Scattered diagram

Figure (A) shows positive correlation between X and Y. Figure (B) shows negative correlation between X and Y.

Figure (C) shows non-linear relationship among the two variables. Figure (D) shows that there is no relationship between the variables.

Algebraic Methods of Correlation

Product moment or covariance method: This method is also called Karl Pearson's method. To measure the intensity or degree of linear relationship between two variables, a formula was developed by Karl Pearson, a famous Biometrician called correlation coefficient.

Correlation coefficient between two random variables X and Y denoted by $r(X, Y)$ or r_{XY} is a numerical measure of linear relation between them which is given by

$$r(X,Y) = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y}$$

Where $\text{cov}(X, Y)$ is the covariance between the two variables X and Y, σ_x and σ_y are the standard deviations of the two variables X and Y, respectively. If the data is (x_i, y_i) then the above terms can be calculated using,

$$\text{cov}(X,Y) = E[\{X - E(X)\}\{Y - E(Y)\}]$$
$$= \frac{1}{n}\sum (x_i - \bar{x})(y_i - \bar{y})$$

$$\sigma_X^2 = E[X - E(X)]^2 = \frac{1}{n}\sum (x_i - \bar{x})^2 \qquad \sigma_Y^2 = E[Y - E(Y)]^2 = \frac{1}{n}\sum (y_i - \bar{y})^2$$

Alternately the following formula can also be used to calculate the above terms:

$$\text{cov}(X,Y) = \frac{1}{n}\sum (x_i - \bar{x})(y_i - \bar{y})$$
$$= \frac{1}{n}(x_i y_i - x_i \bar{y} - \bar{x} y_i + \bar{x}\bar{y})$$
$$= \frac{1}{n}\sum x_i y_i - \bar{y}\frac{1}{n}\sum x_i - \bar{x}\frac{1}{n}\sum y_i + \frac{\sum}{n}(\bar{x}\,\bar{y})$$
$$= \frac{1}{n}\sum x_i y_i - \bar{y}\bar{x} - \bar{x}\bar{y} + \bar{x}\bar{y}$$
$$= \frac{1}{n}\sum x_i y_i - \bar{x}\bar{y}$$
$$\sigma_X^2 = \frac{1}{n}\sum x_i^2 - \bar{x}^2, \quad \sigma_Y^2 = \frac{1}{n}\sum y_i^2 - \bar{y}^2$$

11.3 PROPERTIES OF CORRELATION COEFFICIENT

1. **Limits for correlation coefficient:**

   Let μ_X and μ_Y denote the means of X and Y.
   Let $E(X) = \mu_X$ and $E(Y) = \mu_Y$

$$\text{Consider } E\left[\left(\frac{X-\mu_X}{\sigma_X}\right)\pm\left(\frac{Y-\mu_Y}{\sigma_Y}\right)\right]^2 \geq 0$$

$$E\left[\frac{X-\mu_X}{\sigma_X}\right]^2 + E\left[\frac{Y-\mu_Y}{\sigma_Y}\right]^2 \pm \frac{2E[(X-E(X))((Y-E(Y))]}{\sigma_X\sigma_Y} \geq 0 \tag{11.1}$$

$$\text{but } E[X-\mu_X]^2 = E[X-E(X)]^2$$

$$= \sigma_X^2 \quad \text{and}$$

$$E[Y-\mu_Y]^2 = E[Y-E(Y)]^2$$

$$= \sigma_Y^2$$

$$\therefore E\left[\frac{X-\mu_X}{\sigma_X}\right]^2 = 1,\ E\left[\frac{Y-\mu_Y}{\sigma_Y}\right]^2 = 1$$

Hence equation (11.1) reduces to

$$1+1\pm\frac{2E[(X-E(X))(Y-E(Y))]}{\sigma_X\sigma_Y} \geq 0$$

$$1+1\pm\frac{2\text{cov}(X,Y)}{\sigma_X\sigma_Y} \geq 0,\ 2\pm 2r(X,Y) \geq 0$$

$$-1 \leq r(X,Y) \leq 1.$$

Hence, the correlation coefficient lies between -1 and $+1$. It cannot exceed unity numerically. When $r = +1$, the correlation is perfect and positive and when $r = -1$, the correlation is perfect and negative.

2. Two independent variables are uncorrelated. But the converse of the statement need not be true. That is, if X and Y are independent variables, then $\text{cov}(X, Y) = 0$ and hence $r(X,Y) = \frac{\text{cov}(X,Y)}{\sigma_X\sigma_Y} = 0$

Hence, two independent variables are uncorrelated. However, the converse of the theorem is not true, that is, two uncorrelated variables may not be independent. This is shown in the following example:

Let X be a normal variate and $Y = X^2$.

Since $X \sim N(0, 1)$, $E(X) = 0$

$$E(XY) = E(XX^2) = E(X^3) = 0$$

$$\text{cov}(X, Y) = E(XY) - E(X)E(Y)$$

$$= E(X^3) - E(X)E(X^2)$$

$$= 0$$

$$\therefore r(X,Y) = \frac{\text{cov}(X,Y)}{\sigma_X\sigma_Y} = 0$$

Hence, X and Y are uncorrelated, but they are dependent variables.

3. Correlation coefficient is independent of change of origin and scale.
Let X, Y be two random variables with correlation coefficient between them as

$$r(X,Y) = \frac{\text{cov}(X,Y)}{\sigma_X\sigma_Y}$$

Let us shift the origin and scale to some other point U, V which are defined as:

$$U = \frac{X - a}{h} \text{ and } V = \frac{Y - b}{k}$$

where a, b, h, and k are all constants.

We get $X = a + hU,\ Y = b + kV$

$$E(X) = a + hE(U), \qquad E(Y) = b + kE(V)$$

$$X - E(X) = h[U - E(U)], \qquad Y - E(Y) = k[V - E(V)]$$

$$\begin{aligned}\text{cov}(X, Y) &= E[\{X - E(X)\}\{Y - E(Y)\}] \\ &= E[h\{U - E(U)\}k\{V - E(V)\}] \\ &= hk[\{U - E(U)\}\{V - E(V)\}] = hk\ \text{cov}(U, V)\end{aligned} \tag{11.2}$$

$$\begin{aligned}\sigma_X^2 &= E\{X - E(X)\}^2 = E\{h(U - E(U))\}^2 \\ &= h^2 E\{U - E(U)\}^2 = h^2\sigma_U^2 \Rightarrow \sigma_X = h\sigma_U (h > 0)\end{aligned} \tag{11.3}$$

$$\begin{aligned}\sigma_Y^2 &= E\{Y - E(Y)\}^2 = E\{k(V - E(V))\}^2 \\ &= k^2 E\{V - E(V)\}^2 = k^2\sigma_V^2\end{aligned}$$

$$\sigma_Y = k\sigma_V (k > 0) \tag{11.4}$$

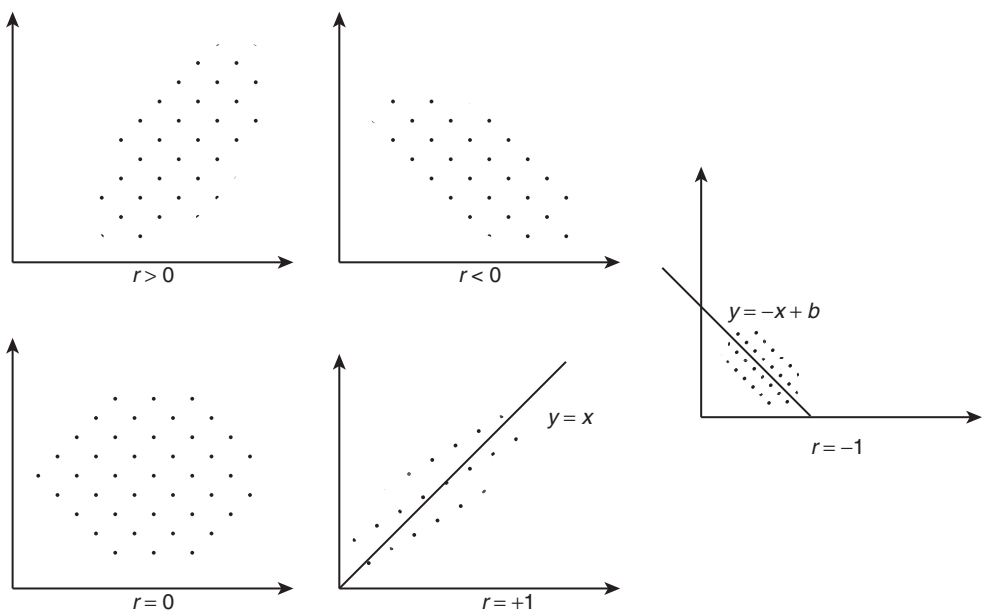
$\therefore$ The correlation coefficient between (X, Y) is

$$r_{XY} = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{hk\ \text{cov}(U, V)}{hk\sigma_U\sigma_V}$$

$$= r_{UV}$$

This property is useful in the numerical computations of correlation coefficient.

4. $r(X, Y)$ provides a linear relationship between X and Y.



5. Karl Pearson correlation coefficient r is based on the following assumptions:
   (i) The variables X and Y are linearly related.
   (ii) Each of the variables or series is being affected by a large number of independent contributory causes of such a nature as to produce normal distribution.
   (iii) The forces so operating on each of the variable series are not independent of each other but are related in a casual fashion.

Worked Out Examples

EXAMPLE 11.1

Calculate the correlation coefficient between the two series X and Y.

X	25	18	32	21	35	29
Y	16	11	20	15	26	28

Solution: When the value of the variables are high we use a shortcut method (θ) by shifting the origin and scale.

X	Y	$U = X - 25$	$V = Y - 20$	U^2	V^2	UV
25	16	0	–4	0	16	0
18	11	–7	–9	49	81	63
32	20	7	0	49	0	0
21	15	–4	–5	16	25	20
35	26	10	6	100	36	60
29	28	4	8	16	64	32
$\Sigma X = 160$	$\Sigma Y = 116$	$\Sigma U = 10$	$\Sigma V = -4$	$\Sigma U^2 = 230$	$\Sigma V^2 = 222$	$\Sigma UV = 175$

$$\bar{U} = \frac{1}{n}\Sigma U = \frac{1}{6}(10) = 1.67$$

$$\bar{V} = \frac{1}{n}\Sigma V = \frac{-4}{6} = 0.67$$

$$\sigma_U^2 = \frac{1}{n}\Sigma U^2 - (\bar{U})^2 = \frac{1}{6}(230) - (1.67)^2$$

$$= 35.544$$

$$\sigma_V^2 = \frac{1}{n}\Sigma V^2 - (\bar{V})^2 = \frac{1}{6}(222) - (0.67)^2$$

$$= 36.5511$$

$$\text{cov}(U,V) = \frac{1}{n}\sum UV - \bar{U}\bar{V} = \frac{1}{6}(175) - (1.67)(0.67) = 28.047$$

$$r(U,V) = \frac{\text{cov}(U,V)}{\sigma_U \sigma_V}$$

$$= \frac{28.047}{\sqrt{(35.544)(36.5511)}}$$

$$= \frac{28.047}{36.044}$$

$$= 0.778$$

This shows a very high correlation between the two series of variables.

EXAMPLE 11.2

Calculate the correlation coefficient between the heights of father and heights of son from the given data.

Heights of father (X)	64	65	66	67	68	69	70
Heights of sons (Y)	66	67	65	68	70	68	72

Solution:

Heights of father (X)	Heights of sons (Y)	$U = \frac{X-66}{2}$	$V = \frac{Y-67}{2}$	U^2	V^2	UV
64	66	–1	–0.5	1	0.25	0.5
65	67	–0.5	0	0.25	0	0
66	65	0	–1	0	1	0
67	68	0.5	0.5	0.25	0.25	0.25
68	70	1	1.5	1	2.25	1.5
69	68	1.5	0.5	2.25	0.25	0.75
70	72	2	2.5	4	6.25	5.0
Total		3.5	3.5	8.75	10.25	8

$$\Sigma U = 3.5,\ \Sigma V = 3.5,\ \Sigma U^2 = 8.75,\ \Sigma V^2 = 10.25,\ \Sigma UV = 8$$

$$\bar{U} = \frac{\Sigma U}{n} = \frac{3.5}{7} = 0.5 \qquad \bar{V} = \frac{\Sigma V}{n} = \frac{3.5}{7} = 0.5$$

$$\sigma_U^2 = \frac{1}{n}\Sigma U^2 - \bar{U}^2 \qquad \sigma_V^2 = \frac{1}{n}\Sigma V^2 - \bar{V}^2$$

$$= \frac{1}{7}(8.75) - (0.5)^2 \qquad = \frac{1}{7}(10.25) - (0.5)^2$$

$$= 1 \qquad = 1.214$$

$$\text{cov}(U,V) = \frac{1}{n}\sum UV - \bar{U}\bar{V}$$
$$= \frac{1}{7}(8) - (0.5)(0.5)$$
$$= 0.8928$$
$$r(U,V) = \frac{\text{cov}(U,V)}{\sigma_U \sigma_V}$$
$$= \frac{0.8928}{\sqrt{1(1.214)}} = \frac{0.8928}{1.1018}$$
$$r(U,V) = 0.8103$$

Since r is independent of change of origin & scale, $r(U,V) = r(x, y) = 0.8103$

EXAMPLE 11.3

If $n = 50$, $\Sigma X = 75$, $\Sigma Y = 80$, $\Sigma X^2 = 130$, $\Sigma Y^2 = 140$, $\Sigma XY = 120$, find the value of r.

Solution:

$$r = \frac{\text{cov}(X,Y)}{\sqrt{\sigma_X^2 \sigma_Y^2}}$$
$$r = \frac{\frac{1}{n}\sum XY - \left(\frac{\sum X}{n}\right)\left(\frac{\sum Y}{n}\right)}{\sqrt{\left[\frac{1}{n}\sum X^2 - \left(\frac{\sum X}{n}\right)^2\right]\left[\frac{1}{n}\sum Y^2 - \left(\frac{\sum Y}{n}\right)^2\right]}}$$
$$= \frac{\frac{1}{50}(120) - \left(\frac{75}{50}\right)\left(\frac{80}{50}\right)}{\sqrt{\frac{1}{50}(130) - \left(\frac{75}{50}\right)^2}\sqrt{\frac{1}{50}(140) - \left(\frac{80}{50}\right)^2}}$$
$$= \frac{2.4 - (1.5)(1.6)}{\sqrt{2.6 - (1.5)^2}\sqrt{2.8 - (1.6)^2}}$$
$$= \frac{2.4 - 2.4}{\sqrt{(0.35)(0.24)}} = 0$$

EXAMPLE 11.4

Calculate the coefficient of correlation between X and Y series from the following data:

	X Series	Y Series
No. of items	15	15
Arithmetic mean	25	18
Sum of squares of deviations from mean	136	138

Sum of the product of deviations of X and Y from respective arithmetic means = 122.

Solution:

$$n = 15,\ \bar{x} = 25,\ \bar{y} = 18,$$

$$\Sigma(x - \bar{x})^2 = 136 \qquad \Sigma(y - \bar{y})^2 = 138,$$

$$\sigma_X^2 = \frac{1}{n}\Sigma(x - \bar{x})^2 = \frac{136}{15} = 9.066$$

$$\sigma_Y^2 = \frac{1}{n}\Sigma(y - \bar{y})^2 = \frac{138}{15} = 9.2$$

$$\text{cov}(X, Y) = \frac{1}{n}\Sigma(x_i - \bar{x})(y_i - \bar{y}) = \frac{(122)}{15} = 8.133$$

Correlation coefficient between X and Y is

$$r(X, Y) = \frac{\text{cov}(X, Y)}{\sqrt{\sigma_x^2 \sigma_Y^2}} = \frac{8.133}{\sqrt{(9.066)(9.2)}}$$

$$= \frac{8.133}{9.13275}$$

$$= 0.8905.$$

EXAMPLE 11.5

The marks obtained by 10 students in mathematics and statistics are given. Find the coefficient of correlation between the two subjects.

Marks in mathematics	75	30	60	80	53	35	15	40	38	48
Marks in statistics	85	45	54	91	58	63	35	43	45	44

Solution:

Marks in maths(X)	Marks in stats(Y)	$U = \frac{X - 35}{10}$	$V = \frac{Y - 54}{10}$	U^2	V^2	UV
75	85	4	3.1	16	9.61	12.4
30	45	–0.5	–0.9	0.25	0.81	0.45
60	54	2.5	0	6.25	0	0
80	91	4.5	3.7	20.25	13.69	16.65
53	58	1.8	0.4	3.24	0.16	0.72
35	63	0	0.9	0	0.81	0

15	35	–2	–1.9	4	3.61	3.8
40	43	2.5	–1.1	6.25	1.21	–2.75
38	45	0.3	–0.9	0.09	0.81	–0.27
48	44	1.3	–1	1.69	1	–1.3
		14.4	2.3	58.02	31.71	29.7

$$\bar{U} = \frac{1}{n}\sum U = \frac{1}{10}(14.4) = 1.44 \qquad \bar{V} = \frac{1}{n}\sum V = \frac{1}{10}(2.3) = 0.23$$

$$\sigma_U^2 = \frac{1}{n}\sum U^2 - \bar{U}^2 = \frac{1}{10}(58.02)(1.44)^2 = 3.7284$$

$$\sigma_V^2 = \frac{1}{n}\sum V^2 - \bar{V}^2 = \frac{1}{10}(31.71)(0.23)^2 = 3.1181$$

$$\text{cov}(U,V) = \frac{1}{n}\sum UV = \bar{U}\bar{V}$$

$$= \frac{1}{10}(29.7) - (1.44)(0.23) = 2.97 - 0.3312$$

$$= 2.6388$$

$$r(U,V) = \frac{\text{cov}(U,V)}{\sqrt{\sigma_U^2\sigma_V^2}} = \frac{2.6388}{\sqrt{(3.7284)(3.1181)}}$$

$$= \frac{2.6388}{3.4096} = 0.7739$$

Since correlation coefficient is independent of change of origin and scale, $r(U, V) = r(X, Y) = 0.77$.

11.4 COEFFICIENT OF CORRELATION FOR GROUPED DATA

Calculation of correlation coefficient for a bivariate frequency distribution:

In the earlier topic we have calculated correlation coefficient for bivariate data but when the data is considerably small. However, when the data is very large then they can be grouped to a two-way table. If the *x*-series has m classes and *y*-series has *n* classes, then these would be $m \times n$ cells for the two-way table, each of which frequencies can be found. The whole set of cell frequencies will define a bivariate frequency distribution.

Consider the following correlation table:

y series ↓ \ x series →		Classes: Mid points x_1 x_2 ... x_i ... x_m	Total of frequencies of y $g(y)$
Classes	y_1 y_2 . . y_j . . y_n	$f(x, y)$	$g(y) = \sum_x f(x, y)$
Total frequencies of x $f(x)$		$f(x) = \sum_y f(x, y)$	$N = \sum_x \sum_y f(x, y)$ $= \sum_y \sum_x f(x, y)$

Here x has m classes and Y has n classes. $f(x, y)$ denotes the cell frequencies given by f_{ij} which denotes the frequency of individuals in the $(i, j)^{th}$ cell.

The marginal distributions of X and Y are calculated as $f(x) = \sum_y f(x, y)$ and $g(y) = \sum_x f(x, y)$ which gives the sum of frequencies along any row and any column.

The grand general N gives the sum of all frequencies.

$$\sum_x \sum_y f(x, y) = \sum_y \sum_x f(x, y) = \sum_x f(x) = \sum_y g(y) = N$$

To calculate the coefficient of correlation:

$$\bar{x} = \frac{1}{N} \sum_x x f(x) \quad \bar{y} = \frac{1}{N} \sum_y y\, g(y)$$

$$\sigma_X^2 = \frac{1}{N} \sum_x \sum_y x^2 f(x, y) - \bar{x}^2 = \frac{1}{N} \sum_x x^2 f(x) - \bar{x}^2$$

$$\text{Similarly, } \sigma_Y^2 = \frac{1}{N} \sum_y y^2 g(y) - \bar{y}^2$$

Worked Out Examples

EXAMPLE 11.6

Find the correlation coefficient between age and salary of 50 workers in a factory.

Age (in years)	Daily pay in (rupees)				
	160–169	170–179	180–189	190–199	200–209
20–30	5	3	1	–	–
30–40	2	6	2	1	–
40–50	1	2	4	2	2
50–60	–	1	3	6	2
60–70	–	–	1	1	5

			$U=\frac{x-13.5}{10}$	−2	−1	0	1	2				
			$x=$	164.5	174.5	184.5	194.5	204.5				
$V=\frac{y-45}{10}$	Mid value y	Age \ Daily pay		160–169	170–179	180–189	190–199	200–209	Total $f(V)$	$Vf(V)$	$V^2f(V)$	$\sum_u UVf(U,V)$
−2	25	20–30		5 (20)	3 (3)	1 (0)	- (0)	- (0)	9	−18	36	23
−1	35	30–40		2 (4)	6 (6)	2 (0)	1 (−1)	- (0)	11	−11	11	9
0	45	40–50		1 (0)	2 (0)	4 (0)	2 (0)	2 (0)	11	0	0	0
1	55	50–60		- (0)	1 (−1)	3 (0)	6 (0)	2 (4)	12	12	12	9
2	65	60–70		- (0)	- (0)	1 (0)	1 (2)	5 (20)	7	14	28	22
		Total $f(U)$		8	12	11	10	9	[50]	[−3]	[87]	[63]
		$uf(U)$		−16	−12	0	10	18	[0]			
		$U^2f(U)$		32	12	0	10	36	[90]			
		$\sum_v UVf(U,V)$		24	8	0	7	24	[63]			

$$\text{Let } U = \frac{X - 184.5}{10},\ V = \frac{Y - 45}{10}$$

$$\bar{U} = \frac{1}{N}\sum Uf(U) = 0$$

$$\bar{V} = \frac{1}{N}\sum Vf(V) = \frac{-3}{50} = -0.06$$

$$\text{cov}(U,V) = \frac{1}{N}\sum_u\sum_v UVf(U,V) - \bar{U}\bar{V}$$

$$= \frac{1}{50}(63) - 0(-0.06)$$

$$= 1.26$$

$$\sigma_U^2 = \frac{1}{N}\sum u^2 f(U) - \bar{U}^2 \qquad \sigma_V^2 = \frac{1}{N}\sum V^2 f(v) - \bar{V}^2$$

$$= \frac{1}{50}(96) - 0 \qquad = \frac{1}{50}(87) - (0.06)^2$$

$$= 1.80 \qquad = 1.7364$$

$$r(U,V) = \frac{\text{cov}(U,V)}{\sigma_U\sigma_V} = \frac{1.26}{\sqrt{(1.80)(1.7364)}} = \frac{1.26}{1.7674}$$

$$r(U, V) = 0.712$$

Since, the correlation coefficient is independent of change of origin and scale,

$$r(U, V) = r(x, y)$$

$$\therefore\ r(x, y) = 0.712$$

EXAMPLE 11.7

The following table gives the distribution of the total population and those who are totally and partially blind among them. Find out if there is any relation between age and blindness:

Age	0–10	10–20	20–30	30–40	40–50	50–60	60–70	70–80
No. of persons (in thousands)	100	60	40	36	24	11	6	3
Blind persons	55	40	40	40	36	22	18	15

Solution: Let us find out the number of blind persons corresponding to 100 thousand, that is, 1 lakh persons in each age group.

The first value would be the same. The second value is obtained as follows:
Out of 60,000 persons, the number of blind persons = 40

$$\therefore \text{Out of 1,00,000 persons, the number of blind persons} = \frac{40}{60,000} \times 1,00,000 = 67$$

Similarly, the third value is obtained as follows:
Out of 40,000 persons, the number of blind persons = 40.

$$\therefore \text{Out of 1,00,000 persons, the number of blind persons} = \frac{40}{40,000} \times 1,00,000 = 100$$

The fourth value is obtained as follows:
Out of 36,000 persons, the number of blind persons = 40

$$\therefore \text{Out of 1,00,000 persons, the number of blind persons} = \frac{40}{36,000} \times 1,00,000 = 111$$

The fifth value is obtained as follows:
Out of 24,000 persons, the number of blind persons = 40

$$\therefore \text{Out of 1,00,000 persons, the number of blind persons} = \frac{36}{24000} \times 1,00,000 = 150$$

The sixth value is obtained as follows:
Out of 11,000 persons, the number of blind persons = 40

$$\therefore \text{Out of 1,00,000 persons, the number of blind persons} = \frac{22}{11000} \times 1,00,000 = 200$$

The seventh value is obtained as follows:
Out of 6,000 persons, the number of blind persons = 40

$$\therefore \text{Out of 1,00,000 persons, the number of blind persons} = \frac{18}{6000} \times 100000 = 300$$

The eighth value is obtained as follows:
Out of 3,000 persons, the number of blind persons = 40

$$\therefore \text{Out of 1,00,000 persons, the number of blind persons} = \frac{15}{3000} \times 100000 = 500$$

Age group X	Mid values x_i	$U = \frac{x_i - 45}{10}$	U^2	Blind/ lakh Y	$V = \frac{y - 150}{100}$	V^2	UV
0–10	5	–4	16	55	–0.95	0.9025	3.80
10–20	15	–3	9	67	–0.83	0.6889	2.4900
20–30	25	–2	4	100	–0.5	0.25	1.0
30–40	35	–1	1	111	–0.39	0.1521	0.390
40–50	45	0	0	150	0.0	0	0
50–60	55	1	1	200	0.5	0.25	0.5
60–70	65	2	4	300	1.5	2.25	3.0
70–80	75	3	9	500	3.5	12.25	10.5
Total	293	–4	44	1483	2.83	16.7435	21.68

The correlation coefficient

$$r(x, y) = r(u, v) = \frac{\text{cov}(u,v)}{\sigma_U \sigma_V}$$

$$\text{cov}(u,v) = \frac{1}{n}\sum uv - \bar{U}\bar{V}$$

$$= \frac{1}{8}(-4)(2.83) - (-0.5)(0.354)$$

$$\bar{U} = \frac{1}{n}\sum u_i = \frac{-4}{8} = -0.5$$

$$\bar{V} = \frac{1}{n}\sum v_i = \frac{+2.83}{8} = 0.354$$

$$= -1.415 + 0.1768$$

$$= -1.238$$

$$\sigma_U^2 = \frac{1}{n}\sum U^2 - (\bar{U})^2 = \frac{1}{8}(44) - (-0.5)^2 = 5.25$$

$$\sigma_V^2 = \frac{1}{n}\sum V^2 - \bar{V}^2 = \frac{1}{8}(16.7435) - (0.354)^2 = 1.9676$$

$$r(U,V) = \frac{\text{cov}(U,V)}{\sigma_U \sigma_V} = \frac{-1.238}{\sqrt{(5.25)(1.9676)}} = \frac{-1.238}{3.214}$$

$$r(U, V) = -0.385$$

Since correlation coefficient is independent of change of origin and scale, $r(X, Y) = -0.385$

EXAMPLE 11.8

Find the frequency distribution of (U, V) where $U = \frac{X - 7.5}{2.5}$, $V = \frac{Y - 15}{+5}$. Find the correlation coefficient between (X, Y) and (U, V). The frequency distribution of (X, Y) is given as follows:

$X\downarrow$ \ $Y\rightarrow$	5	10
10	30	20
20	20	30

Solution: To find $r(X, Y)$ we have

$$r(X,Y) = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y}$$

$$\text{cov}(X,Y) = \frac{1}{n}\sum XY - \bar{X}\bar{Y}$$

	$U = \frac{x-7.5}{2.5}$	−1	1				
$V = \frac{y-15}{5}$	$x\rightarrow$ / $y\downarrow$	5	10	Total $f(V)$	$Vf(V)$	$V^2f(V)$	$\sum_U UVf(U,V)$
−1	10	30 (30)	20 (−20)	50	−50	50	10
1	20	20 (−20)	30 (30)	50	50	50	10
	Total $f(U)$	50	50	$N = 100$	0	100	20
	$Uf(U)$	−50	50	0			
	$U^2f(U)$	50	50	100			
	$\sum_V UVf(U,V)$	10	10	20			

$$\bar{U} = \frac{1}{N}\sum U f(U) = 0$$

$$\bar{V} = \frac{1}{N}\sum V f(V) = 0$$

$$\begin{aligned}\text{cov}(U,V) &= \frac{1}{N}\sum UV f(U,V) - \bar{U}\bar{V} \\ &= \frac{1}{100}(20) - 0 \\ &= 0.2\end{aligned}$$

$$\begin{aligned}\sigma_U^2 &= \frac{1}{N}\sum U^2 f(U) - \bar{U}^2 \\ &= \frac{1}{100}(100) - 0 \\ &= 1\end{aligned} \qquad \begin{aligned}\sigma_V^2 &= \frac{1}{N}\sum V^2 f(V) - \bar{V}^2 \\ &= \frac{1}{100}(100) - 0 \\ &= 1\end{aligned}$$

$$r(U,V) = \frac{\text{cov}(U,V)}{\sqrt{\sigma_U^2\sigma_V^2}} = \frac{0.2}{1} = 0.2$$

Since r is independent of change of origin and scale $r(X, Y) = r(U, V) = 0.2$.

EXAMPLE 11.9

Consider the following probability distribution:

$X\downarrow$ \ $Y\rightarrow$	0	1	2
0	0.1	0.2	0.1
1	0.2	0.3	0.1

Calculate $E(X)$, $V(X)$, Cov(X, Y), and $r(X, Y)$.

Solution:

$x\downarrow$ \ $y\rightarrow$	0	1	2	$p(x)$	$xp(x)$	$x^2p(x)$	$xyp(x,y)$
0	0.1 (0)	0.2 (0)	0.1 (0)	0.4	0.0	0	0
1	0.2 (0)	0.3 (0.2)	0.1 (0)	0.6	0.6	0.5	0.5
$p(x)$	0.3	0.5	0.2	1.0	[0.6]	[0.5]	[0.5]
$yp(x)$	0	0.5	0.4	[0.9]			
$y^2p(x)$	0	0.5	1.6	[2.1]			
$xyp(x,y)$	0	0.3	0.2	[0.5]			

$$\text{cov}(x, y) = \frac{1}{N}\sum xyP(x, y) - E(X)E(Y)$$

$$E(X) = \frac{1}{N}\sum xP(x) = \frac{1}{1}(0.6)$$

$$E(Y) = \frac{1}{N}\sum yP(y) = \frac{1}{1}(0.9)$$

$$E(XY) = \frac{1}{N}\sum xyP(x, y) = (0.5)$$

$$\text{cov}(x, y) = 0.5 - (0.6)(0.9)$$

$$= -0.4$$

$$\sigma_x^2 = \frac{1}{n}\sum x^2P(x) - [E(X)]^2 = 0.5 - (0.6)^2 = 0.14$$

$$\sigma_y^2 = \frac{1}{n}\sum y^2P(y) - [E(Y)]^2 = 2.1 - (0.9)^2 = 1.29$$

$$r(x, y) = \frac{-0.4}{\sqrt{(0.14)(1.29)}} = \frac{-0.4}{0.4249} = -0.9413$$

11.5 RANK CORRELATION

Sometimes we come across some statistical data that are ranked according to size. Whenever the data is given in terms of rank we cannot use Karl Pearson's correlation coefficient, but we have another correlation coefficient called Spearman's correlation coefficient.

Derivation of $\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$:

Let (x_1, y_1), (x_2, y_2), … (x_n, y_n) denote the ranks of n individuals in two characteristics A and B, respectively. Spearman's Rank correlation is the Karl Pearson's correlation coefficient between the ranks treating them as values of variables.

$x_1, x_2, \ldots x_n$ are only the numbers 1, 2, … n arranged in a different order.

$$\sum x = x_1 + x_2 + \ldots + x_n = 1 + 2 + 3 + \ldots + \frac{n(n+1)}{2}$$

$$\sum x^2 = x_1^2 + x_2^2 + \ldots + x_n^2 = 1^2 + 2^2 + \ldots + n^2 = \frac{n(n+1)(2n+1)}{6}$$

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

$$\sigma_x^2 = \frac{\sum x^2}{n} - \bar{x}^2 = \frac{\not{n}(n+1)(2n+1)}{6\not{n}} - \left(\frac{n+1}{2}\right)^2$$

$$= \frac{(n+1)4n + 2 - 3n - 3}{12} = \frac{n^2-1}{12}$$

Since y series also has the same rankings, we get

$$\sigma_y^2 = \frac{n^2-1}{12}\Bigg|,\ \bar{x} = \bar{y} = \frac{n+1}{2},\ \sigma_x^2 = \sigma_y^2 = \frac{n^2-1}{12}$$

Let $d_i = x_i - y_i$ denotes the differences between the ranks of i^{th} individual in the two characteristics.

$$\because \bar{x} = \bar{y},\ d_i = (x_i - \bar{x}) - (y_i - \bar{y})$$

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

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

$$= \sigma_x^2 + \sigma_y^2 - 2\,\mathrm{cov}(x, y) = \frac{n^2 - 1}{12} + \frac{n^2 - 1}{12} - 2\,\mathrm{cov}(x, y)$$

$$\mathrm{cov}(x, y) = \frac{n^2 - 1}{12} - \frac{\sum d_i^2}{2n}$$

Karl Pearson's correlation coefficient:

$$r(X, Y) = \frac{\mathrm{cov}(x, y)}{\sigma_x \sigma_y} = \frac{\frac{n^2 - 1}{12} - \frac{\sum d_i^2}{2n}}{\sqrt{\frac{n^2 - 1}{12} \cdot \frac{n^2 - 1}{12}}} = 1 - \frac{6\sum d_i^2}{n(n^2 - 1)} = \rho(X, Y)$$

Caution:

- This Spearman's correlation can be used when the actual ranks are given or when the ranks are not given.
- However, Pearson's correlation coefficient can be used only when the data is given, not the ranks of the two series.

11.6 LIMITATIONS OF SPEARMAN'S CORRELATION COEFFICIENT METHOD

The limitations of Spearman's correlation coefficient method are as follows:

(i) When the data is a grouped frequency distribution, this Rank correlation method cannot be applied.

(ii) Hence, when the distribution or the number of observations is quite large, and the ranks are not given, then one has to assign ranks to the observations in each of the series, which is a very tedious process.

(iii) This Rank correlation coefficient is a distribution free or non-parametric measure of correlation. Hence, the result may not be as dependable as in the case of ordinary correlation.

Worked Out Examples

EXAMPLE 11.10

The director of a management training programme is interested to know whether there is a positive association between a trainee's score price and his/her joining the programme and the same trainee's score after the completion of the training. The data on 10 trainees is as follows:

Trainee	1	2	3	4	5	6	7	8	9	10
Rank score 1	1	4	10	8	5	7	3	2	6	9
Rank score 2	2	3	9	10	3	6	1	6	7	8

Determine the degree if association between pre-training and post-training scores.

Solution:

Trainee	Rank score I (x_i)	Rank score II (y_i)	Rank difference $d_i = x_i - y_i$	d_i^2
1	1	2	–1	1
2	4	3	1	1
3	10	9	1	1
4	1	10	–2	4
5	5	5	0	0
6	7	6	1	1
7	3	1	2	4
8	2	4	–2	4
9	6	7	–1	1
10	9	8	1	1
			$\sum d_i = 0$	$\sum d_i^2 = 18$

Spearman's Rank correlation is given by

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

$$= 1 - \frac{6(18)}{10(10^2-1)} = 1 - \frac{108}{990}$$

$$= 1 - 0.11 = 0.89$$

Hence, there is a high degree of association between pre-training and post training scores.

EXAMPLE 11.11

Two persons were asked to watch 10 specified TV programmes and offer their evaluation by rating them 1 to 10. These ratings are given as follows:

TV Programme	A	B	C	D	E	F	G	H	I	J
Ranks given by *X*	4	6	3	9	1	5	2	7	10	8
Ranks given by *Y*	2	3	4	9	5	7	1	10	8	6

Calculate Spearman's correlation coefficient of the two ratings.

Solution:

TV Programme	Rank given by $X(x_i)$	Rank given by $Y(y_i)$	Rank difference $d_i = x_i - y_i$	d_i^2
A	4	2	2	4
B	6	3	3	9
C	3	4	–1	1
D	9	9	0	0
E	1	5	–4	16
F	5	7	–2	4
G	2	1	1	1
H	7	10	–3	9
I	10	8	2	4
J	8	6	2	4
				$\sum d_i^2 = 52$

Spearman's Rank correlation is given by

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

$$= 1 - \frac{6(52)}{10(10^2-1)}$$

$$\rho = 0.685$$

11.7 TIED RANKS

If some of the individuals receive the same rank in a ranking of merit, they are called as tied ranks. Suppose m of individuals are tied then each of those individuals is assigned a same rank, which is the arithmetic mean of their number. Hence in the Rank correlation formula, we add the factor $\frac{m(m^2-1)}{12}$ to $\sum d_i^2$ where m is the number of times an item is repeated. This correction factor is to be added for each repeated value in both X and Y series.

Worked Out Examples

EXAMPLE 11.12

Obtain the Rank correlation coefficient from the following data:

X	68	64	75	50	64	80	75	40	55	64
Y	62	58	68	45	81	60	68	48	50	70

Solution: Calculation of Rank correlation coefficient:

X	Y	Rank X x_i	Rank Y y_i	$d_i = x_i - y_i$	d_i^2
68	62	4	5	–1	1
64	58	6	7	–1	1
75	68	2.5	3.5	–1	1
50	45	9	10	–1	1
64	81	6	1	5	25
80	60	1	6	5	25
75	68	2.5	3.5	–1	1
40	48	10	9	1	1
55	50	8	8	0	0
64	70	6	2	4	16
					$\sum d_i^2 = 72$

In the X-series an observation namely 64 has occurred thrice. Hence its rank will be $\frac{5+6+7}{3} = 6$ each. Corresponding to the rank 6, the correction factor to be added $= \frac{m(m^2-1)}{12} = \frac{3(9-1)}{12} = \frac{24}{12} = 2$

The observation 75 is repeated twice. Hence its rank $= \frac{2+3}{2} = 2.5$

Correction factor to be added $= \frac{m(m^2-1)}{12} = \frac{2(4-1)}{12} = \frac{1}{2}.$

Similarly, in Y-series 68 is repeated twice. Its rank $= \frac{3+4}{2} = 3.5$

Correction factor to be added $= \frac{m(m^2-1)}{12} = \frac{2(4-1)}{12} = \frac{1}{2}.$

$$\therefore \Sigma d_i^2 = 72 + \frac{5}{2} + \frac{1}{2} = 75.0$$

Hence, Rank correlation coefficient

$$\rho = 1 - \frac{6\sum d_i^2}{12} = 1 - \frac{6(75.0)}{12} = 0.545$$

EXAMPLE 11.13

Ten competitors in a beauty contest are ranked by three judges in the following order. Find Rank correlation coefficient to determine which pair of judges has the nearest approach to common likings in beauty.

First judge	1	6	5	10	3	2	4	9	7	8
Second judge	3	5	8	4	7	10	2	1	6	9
Third judge	6	4	9	8	1	2	3	10	5	7

Solution:

Rank of judge I X_i	Rank of judge II Y_i	Rank of judge III Z_i	$d_1 = X_i - Y_i$	$d_2 = X_i - Y_i$	d_1^2	d_2^2	$d_3^2 = (Z_i - X_i)^2$
1	3	6	–2	–3	4	9	25
6	5	4	–1	–1	1	1	4
5	8	9	–3	–1	9	1	16
10	4	8	6	–4	36	16	4
3	7	1	–4	6	16	36	4
2	10	2	–8	8	64	64	0
4	2	3	2	–1	4	1	1
9	1	10	8	–9	64	81	36
7	6	5	1	1	1	1	4
8	9	7	–1	–2	1	4	1
					200	214	95

The Rank correlation between I judge and II judge is

$$\rho_1 = 1 - \frac{6\sum d_1^2}{n(n^2-1)} = 1 - \frac{6(200)}{10(100-1)} = -0.212$$

The Rank correlation between judges I and III is

$$\rho_2 = 1 - \frac{6\sum d_2^2}{n(n^2-1)} = 1 - \frac{6(214)}{10(100-1)} = -0.296$$

The Rank correlation between judges I and III is

$$\rho_3 = 1 - \frac{6(95)}{990} = 0.42$$

Since the Rank correlation between judges I and III is less than that between judges I and II and judges II and III the pair of judges I and III is the best pair to judge about the three beauty contestants.

11.8 CONCURRENT DEVIATIONS METHOD

In the previous two topics, we have found the degree of correlation between two variables. Suppose we are interested only in the nature of correlation and not the degree of correlation between the variables, this method can be used.

To compute the coefficient of correlation by this method, the deviations are indicated by '+' sign in the case of increase from the preceding item and '−' sign in the case of decrease from the preceding item and '=' when there is no rise or fall. Those items which have the same sign in the pair, known as concurrent deviations and are marked *C*. If the change is in the reverse direction, it is indicated by a negative sign. The formula used to calculate coefficient of correlation is $r_0 = \overset{(\pm)}{\sqrt{\pm\left(\frac{2C-N}{N}\right)}}$.

where r_0 is the coefficient of concurrent deviations, *C* is the number of concurrent deviations, and *N* is the number of pairs of deviations compared.

Caution:

The significance of ± signs (both inside and outside the square root is that if we get $\frac{2C-N}{N}$ as negative, this is to be multiplied with a '−' sign inside which would make it positive under the square root and also consider '−' sign outside. If $\frac{2C-N}{N}$ is positive, consider + sign inside and outside the square root.

Limitations

(i) The method does not differentiate between small and big changes, that is, if *X* changes from 100 to 101, we put '+' sign similarly if *Y* changes from 68 to 168 we put '+' sign. Hence, both get same sign as they vary in the same direction.

(ii) The value obtained for r_0 only indicates the presence or absence of correlation but not the exact value.

EXAMPLE 11.14

Obtain the coefficient of correlation from the following data using method of concurrent deviations:

Test number	1	2	3	4	5	6	7	8	9	10	11
Marks in statistics	65	40	35	75	63	80	35	20	80	60	50
Marks in Accountancy	60	55	50	56	30	70	40	35	80	75	80

Solution:

Test	Marks in statistics	Deviations from the preceding test	Marks in accountancy	Deviations from the preceding test	Concurrence	Disagreement
1	65		60	−	+	
2	40	−	55	−	+	
3	35	−	50	+	+	
4	75	+	56	−	+	
5	63	−	30	+	+	
6	80	+	70	−		−

7	35	+	40	−	+		
8	20	−	35	+	+		
9	80	+	80	−	+		
10	60	−	75	+		−	
11	50	−	80				
Total					8	2	

Number of pairs of deviations = 10, Number of concurrent deviations, $C = 8$

Coefficient of concurrent deviations is given by

$$r_0 = \pm\sqrt{\pm\left(\frac{2C-N}{N}\right)} = \pm\sqrt{\pm\left(\frac{16-10}{10}\right)}$$

Since $2C - N = +6$, we take positive sign inside and outside the square root.

Therefore, $r_0 = 0.774$

Hence the coefficient of correlation between marks in statistics and marks in accountancy is +0.774 which indicates a high positive correlation between the two subjects.

Work Book Exercises

1. The mileage (Y) can be obtained from a certain gasoline depends on the amount(X) of a certain chemical in the gasoline. Find the coefficient of correlation for the ten observations for X and Y are as follows:

Amount, X	0.10	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Mileage, Y	10.98	11.14	13.17	13.34	14.39	14.63	13.66	13.78	15.43	18.37

[**Ans.:** 0.90]

2. Find coefficient of correlation for the following data:

Fertilizer used (metric tonnes)	15	18	20	24	30	35	40	50
Productivity (metric tonnes)	85	93	95	105	120	130	150	160

[**Ans.:** 0.9917]

3. The following table gives the distribution of items of production and also the relatively defective items among them, according to size groups. Find the coefficient of correlation between size and defect in quality.

Size of group	15–16	16–17	17–18	18–19	19–20	20–21
No. of items	200	270	340	360	400	300
No. of defective items	150	162	170	180	180	120

[**Ans.:** 0.94]

4. Obtain the coefficient of correlation between the sales and expenses of the following 10 firms:

Firms	1	2	3	4	5	6	7	8	9	10
Sales (in thousand ₹)	50	50	55	60	65	65	65	60	60	50
Expenses (in thousand ₹)	11	13	14	16	16	15	15	14	13	13

[**Ans.:** 0.786]

5. Calculate the coefficient of correlation for the following bivariate data by taking $U = \frac{X - 500}{200}$ $V = \frac{Y - 1250}{500}$.

Area under wheat	Total area				
	0–500	500–1000	1000–1500	1500–2000	2000–2500
0–200	12	06	–	–	–
200–400	02	18	04	02	01
400–600	–	04	07	03	–
600–800	–	01	–	02	01
800–1000	–	–	01	02	03

6. The following frequency distribution relates to the age and salary of 100 employees working in an organization. Find the coefficient of correlation.

Age (yrs)	Salary(₹)			
	1300–1400	1400–1500	1500–1600	1600–1700
20–30	4	6	5	2
30–40	2	5	8	5
40–50	8	12	20	2
50–60	–	8	12	1

7. Determine the coefficient of correlation between the security prices and the amounts of annual dividend on the basis of the following data:

Security prices (₹) \ Annual divided	4–8	8–12	12–16	16–20
120–140	–	–	2	4
100–120	–	1	2	3
80–100	–	2	3	–
60–80	2	2	2	–
40–60	4	2	1	–

8. Obtain Spearman's Rank correlation coefficient between the two kinds of assessment of post graduate students' performance in a college.

Students' name	A	B	C	D	E	F	G	H	I
Internal assessment (out of 100)	51	63	73	46	50	60	47	36	60
External assessment (out of 100)	49	72	74	44	58	66	50	30	35

9. The rankings of 10 trainees at the beginning and at the end of a certain course are given. Find Spearman coefficient.

Trainees	A	B	C	D	E	F	G	H	I	J
Ranks before the training	1	6	3	9	5	2	7	10	8	4
Ranks after the training	6	8	3	7	2	1	5	9	4	10

[**Ans.:** 0.394]

10. Find the Rank correlation coefficient from the following data:

X	40	42	48	35	38	40	45	70	55	51
Y	125	124	120	120	130	128	122	110	116	118

11. The following data relates to volume of sales of a fruit item during 8 successive weeks in a market:

Week No.	1	2	3	4	5	6	7	8
Price/Kg (₹)	5.2	6.3	6.8	6.5	5.8	5.4	6.0	6.5
Sales(Qtls)	19.4	17.0	15.1	16.2	17.8	19.5	16.5	15.2

Plot the data in scatter diagram. In addition, compute the coefficient of correlation between price and volume of sales.

12. Obtain
    (i) Karl Pearson's correlation coefficient
    (ii) Spearman's correlation coefficient
    (iii) Go through concurrent deviation method for the following data:

X	56	54	58	55	60	48	52	49	57	53
Y	105	110	104	105	100	115	108	112	116	110

DEFINITIONS AT A GLANCE

Correlation: If a change in one variable brings about a change in the other variable, the two variables are correlated.

Positive and Negative Correlation: If the two variables under study deviate in the same direction, correlation is positive and if they deviate in the negative direction, correlation is negative.

Simple and Multiple Correlations: If the correlation is between only two variables, then it is simple, and if the correlation is between more than two variables, then it is multiple.

Partial and Total Correlation: If the correlation between two variables is calculated by excluding other variables then it is said to be partial correlation. If it is included between all the variables under considerations, then it is total correlation.

Linear and Non-linear Correlation: If in the correlation the variables under study bear a constant ratio, then it is linear. If the amount of change in one variable does not bear a constant ratio, then it is non-linear.

Tied Ranks: When the ranks in a series are repeated then the ranks are tied.

FORMULAE AT A GLANCE

- Karl Pearson's coefficient of correlation is

$$r(X,Y) = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y}$$

where $\text{cov}(X, Y) = E(XY) - E(X)E(Y)$

$$\text{Cov}(X, Y) = \frac{1}{n}\sum(x_i - \bar{x})(y_i - \bar{y})$$
$$= \frac{1}{n}\sum x_i y_i - \bar{x}\,\bar{y}$$
$$\sigma_X = \sum x_i^2 - \bar{x}^2 = \frac{1}{n}\sum(x_i - \bar{x})^2$$
$$\sigma_Y = \sum y_i^2 - \bar{y}^2 = \frac{1}{n}\sum(y_i - \bar{y})^2$$

- The limits for Karl Pearson's correlation coefficient is $-1 \le r \le 1$

OBJECTIVE TYPE QUESTIONS

1. The correlation between two variables is _____, when the variations in the two variables have a constant ratio.
   (a) Partial correlation (b) linear correlation
   (c) simple correlation (d) none

2. Correlation coefficient $r(X, Y)$ is given by
   (a) $\frac{\sigma_X \sigma_Y}{\text{cov}(X,Y)}$ (b) $\frac{\text{cov}(X,Y)}{\sigma_X{}^2 \sigma_Y{}^2}$
   (c) $\frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y}$ (d) none

3. Limits for correlation coefficient are
   (a) $-1 \le r(X, Y) \le 0$ (b) $-1 \le r(X, Y) \le 1$
   (c) $0 \le r(X, Y) \le 1$ (d) none

4. Correlation coefficient is independent of change of
   (a) only origin (b) only scale
   (c) both origin and scale (d) none

5. If $r(U, V) = 0.2$ where $U = \frac{X - 7.5}{2.5}$, $V = \frac{Y - 15}{5}$, then $r(X, Y) =$ _____
   (a) 0.1 (b) 0.2
   (c) 0.3 (d) none

6. Spearman's correlation coefficient is given by $\rho(x, y) =$ _____
   (a) $1 - \frac{6\sum d_i^2}{n(n^2 - 1)}$ (b) $\frac{6\sum d_i^2}{n(n^2 - 1)}$
   (c) $1 - \frac{6\sum d_i^2}{n(n - 1)}$ (d) none

7. The collection factor to be added for each repeated value in both X and Y series in case of tied ranks is _____
   (a) $\frac{m(m^2 - 1)}{12}$ (b) $\frac{m(m - 1)}{12}$
   (c) $\frac{m^2(m^2 - 1)}{12}$ (d) none

8. If $\sum d_i^2 = 52$ for 10 non repeated observations, then Spearman's correlation coefficient is _____

(a) 0.585
(b) 0.785
(c) 0.685
(d) none

ANSWERS

1. (b) 2. (c) 3. (b) 4. (c) 5. (b) 6. (a) 7. (a) 8. (c)

12 Regression

Prerequisites

Before you start reading this unit, you should:

- Know the equation of a straight line
- Know the basic ideas of graphs

Learning Objectives

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

- Estimate the value of a random variable (the dependent variable) given that the value of an associated variable (the independent variable) is known
- Use the least squares estimating equation to predict future values of the dependent variable
- Learn limitations of regression
- Use the least squares method to calculate the equation of a regression line for a given set of data
- Understand the assumptions under which the regression analysis is carried out
- Evaluate as to how good the regression is?

12.1 REGRESSION

Let us first understand the meaning of the word regression. In Unit 11, we have seen the extent to which two variables are related. In this unit, we shall see how to predict one variable provided we have certain data available with us regarding the two variables.

The term regression is first used as a statistical concept in late eighteenth century. The word 'regress' means moving back to the average.

It is a statistical technique to predict one variable from another variable. The known variable is called as independent variable, the variable we are trying to predict is the dependent variable. For example, there is a relationship between the annual sales of aerosol spray cans and the quantity of fluorocarbons released into the atmosphere every year. That is, the number of aerosol cans sold each year is the independent variable and the quantity of fluorocarbons released annually would be the dependent variable.

Definitions

A few definitions given by administrators and statisticians are as follows:

- One of the most frequently used techniques in economics and business research, is to find a relation between two or more variables that are related casually is regression analysis.

  —Taro Yamane

- Regression analysis attempts to establish the nature of relationships between variables, that is, to study the functional relationship between the variables and thereby provides a mechanism for prediction or forecasting.

  —Ya-lun Chou

- The term regression analysis refers to the methods by which estimates are made from the values of a variable, from knowledge of the values of one or more other variables and to the measurement of errors involved in this estimation process.

—Morris Hamburg

Caution:

In regression analysis, the independent variable is known as regressor or predictor or explanatory variable while the dependent variable is called as regressed or explained variable.

Uses of Regression Analysis

Regression analysis is applied to many fields like economics, business, social sciences, statistics, etc. Some of the main uses of studying regression analysis are as follows:

- Regression analysis provides estimates of values of the dependent variable from the independent variable through regression lines.
- In addition, we can obtain a measure of the error involved in using the regression line as a basis for estimation. For this purpose, the standard error of estimate is calculated. In a scatter diagram, if the points are scattered little, that is, the line fits data closely, good estimates can be made on dependent variable, otherwise the line will not produce accurate estimates of the dependent variable.
- With the help of regression coefficient, we can find correlation coefficient.
- Lines of regression—in a bivariate distribution, if the variables are related such that the points in the scatter diagram will cluster round some line, it is called line of regression.

12.2 LINES OF REGRESSION

Let (x_i, y_i), $i = 1, 2 \ldots n$ represent the bivariate data where Y is the dependent variable depending on the independent variable X.

Let the line of regression of Y on X be

$$Y = a + bX \tag{12.1}$$

According to the principle of least squares, the normal equations for estimating the parameters a and b are

$$\sum_{i=1}^{n} y_i = na + b\sum_{i=1}^{n} x_i \tag{12.2}$$

$$\sum_{i=1}^{n} x_i y_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 \tag{12.3}$$

On dividing equation (12.2) by n we get,

$$\frac{1}{n}\sum_{i=1}^{n} y_i = a + b\frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\bar{y} = a + b\bar{x} \tag{12.4}$$

From equation (12.1) and the above equation, it is clear that $(\bar{x}, \bar{y})$ is a point on (12.1).

Equation $\frac{(12.3)}{n}$ gives

$$\frac{1}{n}\Sigma x_i y_i = a\frac{1}{n}\Sigma x_i + \frac{b}{n}\Sigma x_i^2$$

$$= a\bar{x} + \frac{b}{n}\Sigma x_i^2 \tag{12.5}$$

In addition, we know that

$$\text{Cov}(X,Y) = \mu_{11} = \frac{1}{n}\Sigma x_i y_i - \bar{x}\,\bar{y}$$

$$\Rightarrow \frac{1}{n}\Sigma x_i y_i = \mu_{11} + \bar{x}\,\bar{y} \quad \text{and} \tag{12.6}$$

$$\sigma_X^2 = \frac{1}{n}\Sigma x_i^2 - \bar{x}^2$$

$$\Rightarrow \frac{1}{n}\Sigma x_i^2 = \sigma_X^2 + \bar{x}^2 \tag{12.7}$$

Substituting (12.6) and (12.7) in (12.5) we get

$$\mu_{11} + \bar{x}\,\bar{y} = a\bar{x} + b(\sigma_X^2 + \bar{x}^2)$$

Multiplying (12.4) by $\bar{x}$ we get $\bar{x}\,\bar{y} = a\bar{x} + b\bar{x}^2$

$$\text{(i.e.,)}\ \mu_{11} + a\bar{x} + b\bar{x}^2 = a\bar{x} + b\sigma_X^2 + b\bar{x}^2$$

$$\mu_{11} = b\bar{x}^2$$

$$\therefore b = \frac{\mu_{11}}{\sigma_X^2}$$

Since b is the slope of the regression line of y on x, and since the regression line passes through the point $(\bar{x}, \bar{y})$, its equation is

$$(Y - \bar{y}) = b(X - \bar{x})$$

$$Y - \bar{y} = \frac{\mu_{11}}{\sigma_X^2}(X - \bar{x}) \tag{12.8}$$

However, correlation coefficient between x and y is

$$r(X,Y) = r = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{\mu_{11}}{\sigma_X \sigma_Y} = \left(\frac{\mu_{11}}{\sigma_X^2}\right)\frac{\sigma_X}{\sigma_Y}$$

$$\therefore \frac{\mu_{11}}{\sigma_X^2} = r\frac{\sigma_Y}{\sigma_X}$$

Substituting this equation (12.8) we get

$$Y - \bar{y} = r\frac{\sigma_Y}{\sigma_X}(X - \bar{x}) \tag{12.9}$$

This is equation of line of regression of Y on X. Similarly, the equation of line of regression of X on Y is given by

$$X - \bar{x} = r\frac{\sigma_X}{\sigma_Y}(Y - \bar{y}) \tag{12.10}$$

Caution:

We can observe that both the lines of regression, that is, line of regression of Y on X and line of regression of X on Y pass through the point $(\bar{x}, \bar{y})$, the mean values of the data given.

Reason for Two Lines of Regression

Equations (12.9) and (12.10) are two lines of regression, (12.9) being the line of regression of Y on X and (12.10) the line of regression of X on Y. Equation (12.9) is used to estimate any y given a value of x which is derived on minimizing the sum of squares of errors of estimates in Y and not in X. Similarly, we use the line of regression of X on Y to estimate or predict X, which is derived on minimizing the sum of squares of errors of estimates in X and not in Y. Here, X is a dependent variable. Hence the two equations are not interchangeable or reversible because of the basis and assumptions for both the equations that are quite different. The regression equation of Y on X is obtained on minimizing the sum of squares of the errors parallel to the y-axis while the regression equation of X on Y is obtained on minimizing the sum of squares of the errors parallel to the x-axis.

Specific Case $r = \pm 1$

In case of perfect correlation, that is, the coefficient of correlation $r = \pm 1$, then

The equation of line of regression of Y on X becomes

$$Y - \bar{y} = \pm \frac{\sigma_Y}{\sigma_X}(X - \bar{x})$$

$$\left(\frac{Y - \bar{y}}{\sigma_Y}\right) = \pm\left(\frac{X - \bar{x}}{\sigma_X}\right) \tag{12.11}$$

The equation of line of regression of X on Y becomes

$$(X - \bar{x}) = \pm \frac{\sigma_X}{\sigma_Y}(Y - \bar{y})$$

$$\frac{X - \bar{x}}{\sigma_X} = \pm\left(\frac{Y - \bar{y}}{\sigma_Y}\right) \tag{12.12}$$

Equations (12.11) and (12.12) are the same. Hence in the case of perfect correlation, the two lines of regression like the line of regression of Y on X and regression equations of X on Y coincide. Hence in general there are two lines of regression, but when $r = \pm 1$, the two lines of regression coincide.

Worked Out Examples

EXAMPLE 12.1

Obtain regression of Y on X and estimate Y when $X = 55$ from the following:

X	40	50	38	60	65	50	35
Y	38	60	55	70	60	48	30

Solution: Calculation of Regression equation:

X	Y	$U = \frac{X - 40}{10}$	$V = \frac{Y - 55}{10}$	U^2	V^2	UV
40	38	0	1.7	0	2.89	0
50	60	1	0.5	1	0.25	0.5

38	55	−0.2	0	0.04	0.0	0
60	70	2	1.5	4	2.25	3
65	60	2.5	0.5	6.25	0.25	1.25
50	48	1	−0.7	1	0.49	−0.7
35	30	−0.5	−2.5	0.25	6.25	1.25
	Total	5.8	1	12.54	12.38	5.3

$$\bar{U} = \frac{\Sigma U}{n} = \frac{5.8}{7} = 0.828, \quad \bar{V} = \frac{\Sigma V}{n} = \frac{1}{7} = 0.143$$

$$\sigma_U^2 = \frac{1}{n}\Sigma U^2 - \bar{U}^2 = \frac{1}{7}(12.54) - (0.828)^2 = 1.7914 - 0.6855 = 1.1059$$

$$\sigma_V^2 = \frac{1}{n}\Sigma V^2 - \bar{V}^2 = \frac{1}{7}(12.38) - (0.143)^2$$
$$= 1.7685 - 0.0204 = 1.7481$$

$$r(U,V) = \frac{\text{cov}(U,V)}{\sigma_U \sigma_V} = \frac{\frac{1}{n}\Sigma UV - \bar{U}\bar{V}}{\sqrt{(1.1059)(1.7481)}} = \frac{\frac{1}{7}(5.3) - (0.828)(0.143)}{1.3904}$$
$$= 0.4593$$

The regression equation of Y on X is

$$Y - \bar{y} = r\frac{\sigma_Y}{\sigma_X}(X - \bar{x}) \Rightarrow Y - 51.57 = (0.4593)\left(\frac{13.22}{10.516}\right)(X - 48.285)$$
$$= 0.57746(X - 48.285)$$

$$Y = 0.5774X + 23.686$$

$$\sigma_X^2 = h^2\sigma_U^2 \qquad \sigma_Y^2 = k^2\sigma_V^2$$
$$= 100(1.1059) \qquad = 100(1.17481)$$
$$= 110.59 \qquad = 174.81$$
$$\sigma_X = 10.516 \qquad \sigma_Y = 13.22$$
$$\bar{x} = \frac{\Sigma X}{n} = 48.285, \qquad \bar{y} = \frac{\Sigma Y}{n} = 51.57$$

The regression equation of X on Y is

$$X - \bar{x} = r\frac{\sigma_X}{\sigma_Y}(Y - \bar{y})$$

$$X - 48.285 = (0.4593)\left(\frac{10.516}{13.22}\right)(Y - 51.57)$$

$$= 0.3653(Y - 51.57)$$
$$X = 0.3653Y + 29.443$$

EXAMPLE 12.2

The following table gives the aptitude test scores and productivity indices of 10 workers selected at random:

Aptitude scores (X)	60	62	65	70	72	48	53	73	65	82
Productivity Index (Y)	68	60	62	80	85	40	52	62	60	81

Calculate the following:

(i) Regression equation of Y on X

(ii) Regression line of X on Y

(iii) The productivity index of a worker whose test score is 92

(iv) The test score of a worker whose productivity index is 75

Solution:

Scores (X)	Productivity index (Y)	$U = \frac{X-60}{10}$	$V = \frac{Y-68}{10}$	U^2	V^2	UV
60	68	0	0	0	0	0
62	60	0.2	−0.8	0.04	0.64	−0.16
65	62	0.5	−0.6	0.25	0.36	−0.3
70	80	1	1.2	1	1.44	1.2
72	85	1.2	1.7	1.44	2.89	2.04
48	40	−1.2	−2.8	1.44	7.84	3.36
53	52	−0.7	−1.6	0.49	2.56	1.12
73	62	1.3	−0.6	1.69	0.36	−0.78
65	60	0.5	−0.8	0.25	0.64	−0.4
82	81	2.2	−1.3	4.84	1.69	−2.86
650	650	5	−5.6	11.44	18.42	3.22

$$\bar{x} = \frac{\Sigma X}{n} = \frac{650}{10} = 65, \quad \bar{y} = \frac{650}{10} = 65$$
$$\bar{u} = \frac{\Sigma U}{n} = \frac{5}{10} = 0.2, \quad \bar{v} = \frac{-5.6}{10} = -0.56$$

$$\sigma_U^2 = \frac{1}{n}\Sigma U^2 - \bar{u}^2 = \frac{1}{10}(11.44) - (0.2)^2 = 1.104$$

$$\sigma_V^2 = \frac{1}{n}\Sigma V^2 - \bar{v}^2 = \frac{1}{10}(18.42) - (-0.56)^2 = 1.5284$$

$$\text{cov}(U,V) = \frac{1}{n}\Sigma UV - \bar{U}\bar{V} = \frac{1}{10}(3.22) + (0.2)(0.56) = 0.434$$

$$R(U,V) = \frac{\text{cov}(U,V)}{\sigma_U \sigma_V} = \frac{0.434}{\sqrt{(1.104)(1.5284)}} = \frac{0.434}{1.2989} = 0.3341$$

(i) The regression equation of Y on X is

$$Y - \bar{y} = r\frac{\sigma_Y}{\sigma_X}(X - \bar{x}) \Rightarrow Y - 65 = (0.3341)\frac{(12.3628)}{10.507}(X - 65)$$
$$= 0.3931(X - 65)$$
$$Y = 0.3931X + 139.44$$

$\sigma_X^2 = h^2\sigma_U^2$	$\sigma_Y^2 = k^2\sigma_V^2$
$= 100(1.104)$	$= 100(1.5284)$
$= 110.4$	$= 152.84$
$\sigma_X = 10.507$	$\sigma_Y = 12.3628$

(ii) The regression line of X on Y is $X - \bar{x} = r\frac{\sigma_X}{\sigma_Y}(Y - \bar{y})$

$$X - 65 = \frac{(0.3341)(10.507)}{12.3628}(Y - 65)$$
$$= 0.2839(Y - 65)$$
$$= 0.2839Y - 65(0.2839)$$
$$X = 0.2839Y + 46.5434$$

(iii) When $X = 92$, $Y = 75.6052$ using the regression line of Y on X. When test score is 92, productivity index = 75.6052.

(iv) When $Y = 75$, $X = 67.8359$ using the regression line of X on Y. When productivity index = 75, test aptitude score = 67.8359.

12.3 REGRESSION COEFFICIENTS

Consider the regression line of Y on X.

$$Y - \bar{y} = b(X - \bar{x}) = \frac{\mu_{11}}{\sigma_X^2}(X - \bar{x})$$
$$= \frac{r\sigma_Y}{\sigma_X}(\bar{x} - \bar{x})$$

b, which is slope of the line of regression of Y on X is called regression coefficient. It represents the increment in the dependent variable Y corresponding to a unit change in the independent variable X.

(i.e.,) b_{yx} = Regression coefficients of Y on X

$$= \frac{\mu_{11}}{\sigma_X^2} = r\frac{\sigma_Y}{\sigma_X}$$

Similarly, regression coefficients of X on Y is

$$b_{XY} = \frac{\mu_{11}}{\sigma_Y^2} = r\frac{\sigma_X}{\sigma_Y}$$

Properties of Regression Coefficients

1. Regression coefficients are independent of change of origin and not of scale.

   Let $U = \frac{X-a}{h}$ $V = \frac{Y-b}{k}$ where a, b, h, and k are constants when that $h > 0$, $k > 0$.

   $$X = a + Uh,\ Y = b + Vk$$

   $$E(X) = hE(U),\ E(Y) = kE(V)$$

   In the Unit 11, we have obtained that

   $$\text{cov}(X,Y) = hk\,\text{cov}(U,V)$$

   $$\sigma_X^2 = h^2\sigma_U^2$$

   $$\sigma_Y^2 = k^2\sigma_V^2$$

   $\therefore b_{YX}$ = regression coefficient of Y on X

   $$= r\frac{\sigma_Y}{\sigma_X} = \frac{\mu_{11}}{\sigma_X^2} = \frac{\text{cov}(X,Y)}{\sigma_X^2}$$

   $$b_{YX} = \frac{hk\,\text{cov}(U,V)}{h^2\sigma_U^2} = \frac{k}{h}\frac{\text{cov}(U,V)}{\sigma_U^2}$$

   $$b_{YX} = \frac{k}{h}b_{VU}$$

   Hence, a regression coefficient is not independent of change of scale.

   Similarly, regression coefficient of X on Y is

   $$b_{XY} = \frac{h}{k}b_{UV}$$

2. Correlation coefficient is the geometric mean between the regression coefficients.

   The regression coefficient of X on Y is

   $$b_{XY} = r\frac{\sigma_X}{\sigma_Y}$$

   The regression coefficient of Y on X is given by

   $$b_{XY} = r\frac{\sigma_Y}{\sigma_X}$$

Multiplying the two regression coefficient, we get

$$b_{XY} \cdot b_{YX} = \left(r \frac{\sigma_X}{\sigma_Y} \right) \left(r \frac{\sigma_Y}{\sigma_X} \right)$$

$$= r^2$$

$$r = \pm \sqrt{b_{XY} \cdot b_{YX}}$$

The sign of r depends on the regression coefficients. If the regression coefficients are positive, r is positive, if the regression coefficients are negative, r is negative.

3. If one of the regression coefficients is greater than unity, the other must be less than unity.
   Let one of the regression coefficients be greater than unity.
   (i.e.,) Let the regression coefficient of Y on X be greater than unity.

$$b_{YX} > 1$$

$$\frac{1}{b_{YX}} < 1$$

However, since the correlation coefficient is always between −1 and +1, $r^2 \leq 1$

$$b_{XY} \cdot b_{YX} \leq 1$$

$$b_{XY} \leq \frac{1}{b_{YX}} < 1$$

$$\therefore b_{XY} < 1$$

$\therefore$ Regression coefficient of X on Y is less than 1.

4. Arithmetic mean of the regression coefficients is greater than correlation coefficient r, provided $r > 0$.
   Arithmetic mean of the regression coefficients

$$= \frac{b_{YX} + b_{XY}}{2}$$

We have to prove that

$$\frac{1}{2}(b_{XY} + b_{YX}) \geq r$$

$$\frac{1}{2}\left[r \frac{\sigma_X}{\sigma_Y} + r \frac{\sigma_Y}{\sigma_X} \right] \geq r$$

$$\frac{1}{2}\left[\frac{\sigma_X^2 + \sigma_Y^2}{\sigma_X \sigma_Y} \right] \geq 1$$

$$\Rightarrow \sigma_X^2 + \sigma_Y^2 \geq 2\sigma_X \sigma_Y$$

$$\sigma_X^2 + \sigma_Y^2 - 2\sigma_X \sigma_Y \geq 0$$

$$(\sigma_X - \sigma_Y)^2 \geq 0$$

This is always true since it is a square of a real quantity which is always greater than zero. Hence our assumption $\frac{1}{2}(b_{YX} + b_{XY}) \geq r$ is correct.

12.4 DIFFERENCE BETWEEN REGRESSION AND CORRELATION ANALYSIS

- The correlation coefficient is a measure of degree of variability between X and Y, regression analysis is to study the nature of relationship between the variables so that we can predict the value of given one variable.
- In correlation analysis, we cannot say that one variable is the cause and the other is the effect. In regression analysis, one variable is taken as dependent and other is independent, thus making it possible to study the cause and effect of relationship.
- In correlation analysis, r_{XY} is a measure of direction and degree of linear relationship between X and Y. Hence $r_{XY} = r_{YX}$, it is immaterial whether X is dependent or Y is dependent. But in regression analysis, $b_{YX} \neq b_{XY}$ and hence it makes a difference as to which variable is dependent and which is independent.
- Correlation coefficient is independent of change of scale and origin, but regression coefficients are independent of change of origin but not scale.

Worked Out Examples

EXAMPLE 12.3

Calculate the regression coefficients of Y on X and X on Y from the following data:
$\sum X = 50$, $\sum Y = 60$, $\bar{x} = 5$, $\bar{y} = 6$, $\sum XY = 350$, $\sigma_X^2 = 4$, $\sigma_Y^2 = 9$.

Solution:

$$\begin{aligned}
\text{cov}(X,Y) &= \frac{1}{n}\Sigma_{XY} - \bar{x}\,\bar{y} \\
&= \frac{1}{10}(350) - 50(60) \\
&= 5 \\
\bar{x} &= \frac{\Sigma_X}{n} \\
5 &= \frac{50}{n} \Rightarrow n = 10 \\
r(X,Y) &= \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{5}{2(6)} = 0.833
\end{aligned}$$

The regression equation of Y on X is

$$\begin{aligned}
Y - \bar{y} = r\frac{\sigma_Y}{\sigma_X}(X - \bar{x}) &\Rightarrow Y - 6 = 0.833\left(\frac{3}{2}\right)(X - 5) \\
&= 1.2495X - 0.2475 \\
Y &= 1.2495X - 0.2475 \\
\therefore b_{YX} &= 1.2495
\end{aligned}$$

The regression equation of X on Y is

$$\begin{aligned}
X - \bar{x} = r\frac{\sigma_X}{\sigma_Y}(Y - \bar{y}) \quad &\Rightarrow X - 5 = 0.833\left(\frac{2}{3}\right)(Y - 6) \\
&= 0.5553(Y - 6)
\end{aligned}$$

$$X = 0.5553Y + 1.668$$
$$b_{XY} = 0.5553$$

$\therefore$ The regression coefficients of Y on X and X on Y are 1.2495 and 0.5553.

EXAMPLE 12.4

The correlation coefficients between two variables X and Y is $r = 0.6$. If $\sigma_X = 1.50$, $\sigma_Y = 2.00$, $\bar{x} = 10$ and $\bar{y} = 20$, find the regression lines of (i) Y on X (ii) X on Y.

Solution:

(i) Regression line of Y on X is

$$Y - \bar{y} = r\frac{\sigma_Y}{\sigma_X}(X - \bar{x})$$
$$Y - 20 = 0.6\left(\frac{2.00}{1.50}\right)(X - 10)$$
$$= 0.8(X - 10)$$
$$= 0.8X - 8$$
$$Y = 0.8X + 12$$

(ii) Regression line of X on Y is

$$X - \bar{x} = r\frac{\sigma_X}{\sigma_Y}(Y - \bar{y})$$
$$X - 10 = 0.6\left(\frac{1.5}{2.0}\right)(Y - 20)$$
$$= 0.45(Y - 20) = 0.45Y - 9$$
$$X = 0.45Y + 1$$

EXAMPLE 12.5

Obtain the coefficient of correlation if the two lines of regression are given by

$$2X = 8 - 3Y \text{ and } 2Y = 5 - X.$$

Solution: The regression line of Y on X is

$$2Y = 5 - X$$
$$Y = -\frac{1}{2}X + \frac{5}{2}$$

The regression coefficient of Y on X is $b_{YX} = -\frac{1}{2}$

The regression line of X on Y is

$$2X = 8 - 3Y$$
$$X = -\frac{3}{2}Y + 4$$

The regression coefficient of X on Y is $b_{XY} = -\frac{3}{2}$

From the properties of regression coefficients

$$b_{YX} \cdot b_{XY} = r^2 \Rightarrow r^2 = \left(-\frac{1}{2}\right)\left(-\frac{3}{2}\right)$$

$$r = -0.866$$

The sign of r is in accordance with b_{XY} and b_{YX}.

EXAMPLE 12.6

Are the following statements true? For a bivariate distribution:

(i) $b_{XY} = 2.8$ and $b_{YX} = -0.3$

(ii) $b_{XY} = -0.8$, $b_{YX} = -1.2$ and $r = 0.98$

Solution:

(i) Given that $b_{XY} = 2.8$ and $b_{YX} = -0.3$

From the properties of regression coefficients, both of them will have the same sign. Since one of them is given to be positive and one of them negative, the information given is not true.

(ii) We are given $b_{XY} = -0.8$, $b_{YX} = -1.2$ and $r = 0.98$.

The two regression coefficients are given to be negative, but correlation coefficients also has the same sign as that of regression coefficients, the value of r has to be $r = -\sqrt{(-0.8)(-1.2)} = -0.98$.

Hence, the value of $r = -0.98$ and not 0.98.

EXAMPLE 12.7

Given below is the information regarding the advertisement expenditure and sales in crores of rupees:

(i) Calculate two regression lines.

(ii) Find the likely sales when advertisement expenditure is ₹25 crores.

(iii) If the company wants to attain sales target of ₹150 crores, what could be the advertisement budget?

	Advertisement Expenditure (X)	Sales (Y)
Mean	20	120
SD	5	25

Correlation coefficient is 0.8.

Solution: Given

$$\bar{x} = 20 \quad \bar{y} = 120$$
$$\sigma_X = 5 \quad \sigma_Y = 25$$
$$r = 0.8$$

(i) The regression line of X on Y is

$$X - \bar{x} = r\frac{\sigma_X}{\sigma_Y}(Y - \bar{y})$$
$$X - 20 = 0.8\left(\frac{5}{25}\right)(Y - 120)$$
$$X - 20 = 0.16(Y - 120) \quad \Rightarrow X = 0.8 + 0.16Y$$

This is regression line of X on Y.

Similarly, the regression line of X on Y is

$$Y - \bar{y} = r\frac{\sigma_X}{\sigma_Y}(X - \bar{x})$$
$$Y - 120 = 0.8\left(\frac{25}{5}\right)(X - 20)$$
$$= 4(X - 20) \quad \Rightarrow Y = 40 + 4X$$

(ii) When advertising expenditure $X = ₹25$ crores

$$\text{Sales, } Y = 40 + Y(25) = 140 \text{ crores.}$$

(iii) When sales target is ₹150 crores, $Y = 150$ crores advertisement budget

$$X = 0.8 + 0.164$$
$$= 0.8 + 0.16(150)$$
$$= 24.8 \text{ crores}$$

EXAMPLE 12.8

Given $\bar{x} = 25,\ \bar{y} = 22,\ \sigma_X = 4,\ \sigma_Y = 5,\ r = 0.42$ estimate the value of X when $Y = 12$.

Solution: The regression equation of X on Y is

$$X - \bar{x} = r\frac{\sigma_X}{\sigma_Y}(Y - \bar{y})$$
$$X - 25 = \left(\frac{4}{5}\right)(0.42)(Y - 22)$$
$$X - 25 = 0.336(Y - 22)$$
$$X = 17.608 + 0.334$$

When $Y = 12$, $X = 17.608 + 0.336(12)$

$= 17.608 + 4.032$

$= 21.64$

12.5 ANGLE BETWEEN TWO LINES OF REGRESSION

Equation of line regression of Y on X is $Y - \bar{y} = r\frac{\sigma_Y}{\sigma_X}(X - \bar{x})$ and

Equation of line of regression of X on Y is

$$X - \bar{x} = r\frac{\sigma_X}{\sigma_Y}(Y - \bar{y}) \quad \Rightarrow Y - \bar{y} = \frac{\sigma_Y}{r\sigma_X}$$

Slopes of these two lines are

$$m_1 = r\frac{\sigma_Y}{\sigma_X} \text{ and } m_2 = \frac{\sigma_Y}{r\sigma_X}$$

If θ is the angle between these two lines of regression then,

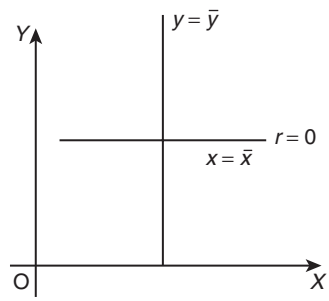
$$\tan\theta = \frac{m_1 \sim m_2}{1 + m_1 m_2} = \frac{r\frac{\sigma_Y}{\sigma_X} \sim \frac{\sigma_Y}{r\sigma_X}}{1 + \left(\frac{r\sigma_Y}{\sigma_X}\right)\left(\frac{\sigma_Y}{r\sigma_X}\right)}$$

$$= \frac{\frac{r^2\sigma_X\sigma_Y \sim \sigma_Y\sigma_X}{r\sigma_X^2}}{\frac{r\sigma_X^2 + r\sigma_Y^2}{r\sigma_X^2}} = \left(\frac{r^2 \sim 1}{r}\right)\left(\frac{\sigma_X\sigma_Y}{\sigma_X^2 + \sigma_Y^2}\right)$$

However, since $r^2 \le 1$, $\tan\theta = \left(\frac{1 - r^2}{r}\right)\left(\frac{\sigma_X\sigma_Y}{\sigma_X^2 + \sigma_Y^2}\right)$

$$\Rightarrow \theta = \tan^{-1}\left[\left(\frac{1 - r^2}{r}\right)\frac{\sigma_X\sigma_Y}{\sigma_X^2 + \sigma_Y^2}\right] \qquad (12.13)$$

From this we can discuss for different values of r, the lines of regression, as follows.



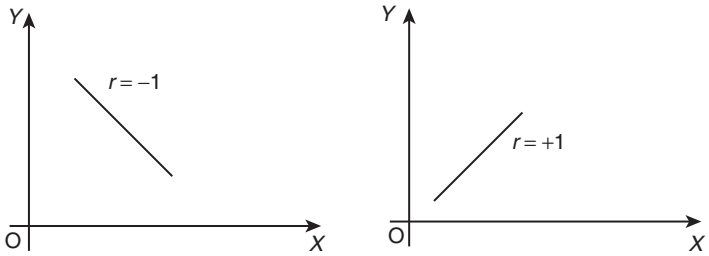
Two lines of regression perpendicular to each other

Case 1: If $r = 0$, from equation (12.13) $\tan\theta = \infty \Rightarrow \theta = \frac{\pi}{2}$

Thus, if two variables are uncorrelated the lines of regression become perpendicular to each other.

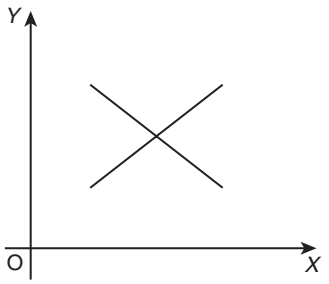
Case 2: If $r = \pm 1$, $\tan \theta = 0 \Rightarrow \theta = 0$ or π

When there is perfect correlation, the two lines of regression either coincide and they are parallel to each other. However, since both the lines of regression pass through the point $(\bar{x}, \bar{y})$, they cannot be parallel. Hence, in case of perfect correlation, the two lines of regression coincide.



Two regression lines coincide

Case 3: When the two lines intersect the angle between them is either acute $\left(0 < \theta < \frac{\pi}{2}\right)$ or obtuse $\left(\frac{\pi}{2} < 0 < \pi\right)$.



Two lines are far away (poor degree of correlation)

Caution:

(i) The higher the degree of correlation between the variables, the angle between the lines is smaller (the two lines are nearer to each other).

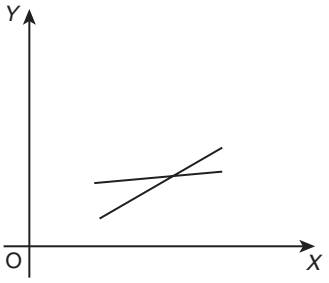
(ii) If the lines of regression make a larger angle, then a poor correlation exists between the variables.

12.6 STANDARD ERROR OF ESTIMATE

So far we have learnt how to estimate one variable with the help of other. Now, we shall learn to measure the reliability of estimating equation that was developed. In a scatter diagram, the line would be more accurate as an estimator if data points lie close to the line, than they are farther away from the line.

The standard error of estimate measures the variability or scatter of the observed values around the regression line.

Consider the regression line of Y on X



Two lines apart (high degree of correlation)

$$Y - \bar{y} = r\frac{\sigma_Y}{\sigma_X}(X - \bar{x})$$

$$\Rightarrow Y = \bar{y} + r\frac{\sigma_Y}{\sigma_X}(X - \bar{x})$$

$$\frac{Y - \bar{y}}{\sigma_Y} = r\left(\frac{X - \bar{x}}{\sigma_X}\right) \qquad (12.14)$$

The standard error of estimates, also called residual variance is the expected value of squares of derivations of the observed values of Y from the expected values.

$$s_Y^2 = E[Y - \hat{y}]^2, \hat{y} \text{ is the estimate value.}$$

From (12.14), $S_Y^2 = E\left[Y - \left\{\bar{y} + r\frac{\sigma_Y}{\sigma_X}(X - \bar{x})\right\}\right]^2$

$$= E\left[(Y - \bar{y}) - r\frac{\sigma_Y}{\sigma_X}(X - \bar{x})\right]^2$$

$$= \sigma_Y^2 E\left[\frac{Y - \bar{y}}{\sigma_Y} - r\left(\frac{X - \bar{x}}{\sigma_X}\right)\right]^2$$

$$= \sigma_Y^2 E[Y^* - rX^*]^2$$

where X^* and Y^* are standardized variates given by $X^* = \frac{X - \bar{x}}{\sigma_X}$ $Y^* = \frac{Y - \bar{y}}{\sigma_Y}$ such that

$$E(X^{*2}) = E\left[\frac{(X - \bar{x})}{\sigma_X}\right]^2$$

$$= E\left[\frac{(X - \bar{x})}{\sigma_X^2}\right] = E\frac{(X - \bar{x})^2}{\sigma_X^2} = 1$$

Similarly, $E(Y^{*2}) =$ and

$$E(X^*Y^*) = E\left[\left(\frac{X - \bar{x}}{\sigma_X}\right)\left(\frac{Y - \bar{y}}{\sigma_Y}\right)\right]$$

$$= E\left[\frac{(X - \bar{x})(Y - \bar{y})}{\sigma_X\sigma_Y}\right] = E\left[\frac{(X - \bar{x})(Y - \bar{y})}{\sigma_X\sigma_Y}\right]$$

$$= r(X, Y) = r$$

$$\therefore S_Y^2 = \sigma_Y^2 E[Y^* - rX^*]^2$$

$$= \sigma_Y^2 [E(Y^{*2}) + r^2 E(X^{*2}) - 2rE(X^*Y^*)]$$

$$S_Y^2 = \sigma_Y^2 [1 + 1 - 2r^2] = 2\sigma_Y^2 (1 - r^2)$$

$$\therefore S_Y = \sigma_Y (1 - r^2)^{\frac{1}{2}}$$

Similarly, the standard error of estimate of X is

$$s_X = \sigma_X (1 - r^2)^{\frac{1}{2}}$$

Caution:

If $r = \pm 1$, $S_X = S_Y = 0$ so that each derivation is zero and the two lines of regression are coincident.

12.7 LIMITATIONS OF REGRESSION ANALYSIS

Limitations for using regression analysis are as follows:

- If one or two extreme values are included, the relationship between the variables may change completely.
- To know whether a linear or nonlinear relationship exists, it is advisable to draw a scatter diagram.

12.8 REGRESSION CURVES

So far we have learned about regression lines, some of their properties, and computations. Now we shall deal with regression curves.

The conditional mean $E\left(\frac{Y}{X}\right) = x$ for a continuous distribution is called regression function of Y on X and its graph is called regression curve Y on X. If the regression curve is a straight line, then the corresponding regression is linear. If one of the regressions is linear, then the other need not be linear.

The following result gives the condition for the regression curve to be linear:

RESULT-1: Let X and Y be two continuous random variables where (X, Y) is a two dimensional random variable. Let $E(X) = \bar{x}$, $E(Y) = \bar{y}$, $\sigma_X^2 = V(X)$, $\sigma_Y^2 = V(Y)$ and $r(X, Y) = r$ is the correlation coefficient between X and Y, then if the regression of Y on X is linear then $E\left(\frac{Y}{X}\right) = \bar{Y} + r\frac{\sigma_Y}{\sigma_X}(X - \bar{X})$. Similarly, if the regression of X on Y is linear, then $E\left(\frac{X}{Y}\right) = \bar{X} + r\frac{\sigma_X}{\sigma_Y}(Y - \bar{Y})$.

Solution: Let the regression equation of Y on X be

$$E\left(\frac{Y}{X}\right) = a + bx \tag{12.15}$$

By definition of conditional mean we have

$$E\left(\frac{Y}{X}\right) = \int_{-\infty}^{\infty} y\, f(y(x))\, dy$$

$$= \int_{-\infty}^{\infty} y \frac{f(x,y)}{f_X(x)} dy$$

where $f_X(x)$ denotes the marginal density of X and $f(x, y)$ denotes the joint probability density functions of x and y.

$$\therefore \frac{1}{f_X(x)} \int_{-\infty}^{\infty} y\, f(x,y) dy = a + bx \tag{12.16}$$

$$\int_{-\infty}^{\infty} y\, f(x,y) dy = (a + bx) f_X(x)$$

$$= a f_X(x) + bx f_X(x)$$

Integrating both sides w. r. t. x between $-\infty$ and ∞,

$$\int_{-\infty}^{\infty}\left[\int_{-\infty}^{\infty} y\, f(x,y)dy\right]dx = a\int_{-\infty}^{\infty} f_X(x)dx + b\int_{-\infty}^{\infty} x\, f_X(x)dx$$

$$\int_{-\infty}^{\infty} y\left[\int_{-\infty}^{\infty} f(x,y)dx\right]dy = a + bE(X)$$

$$\int_{-\infty}^{\infty} y\, f_Y(y)dy = a + bE(X)$$

$$E(Y) = a + bE(X)$$

$$\bar{Y} = a + b\bar{X} \tag{12.17}$$

Multiplying (12.16) with $x\, f_X(x)$ we get

$$\int_{-\infty}^{\infty} x\, yf(x,y)dy = a\, xf_X(x) + bx^2 f_X(x)$$

Integrating both sides w. r. t. x between $-\infty$ and $+\infty$

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy\, f(x,y)dxdy = a\int_{-\infty}^{\infty} x\, f_X(x)dx + b\int_{-\infty}^{\infty} x^2 f_X(x)dx$$

$$E(XY) = a\, E(X) + b\, E(X^2) \tag{12.18}$$

We know that

$$\mu_{11} = E(XY) - E(X)E(Y)$$
$$= E(XY) - \bar{X}\bar{Y}$$
$$\sigma_X^2 = E(X^2) - [E(X)]^2 = E(X^2) - \bar{X}^2$$
$$\sigma_Y^2 = E(Y^2) - \bar{Y}^2$$

$$\therefore E(XY) = \mu_{11} + \bar{X}\bar{Y}$$

From (12.18), $\mu_{11} + \bar{X}\bar{Y} = a\bar{X} + b(\sigma_X^2 + \bar{X}^2)$

From (12.17), $\mu_{11} + \bar{X}\bar{Y} = \bar{X}\bar{Y} - b\bar{X}^2 + b\sigma_X^2 + b\bar{X}^2$

$$\mu_{11} = b\bar{X}^2 \Rightarrow b = \frac{\mu_{11}}{\bar{X}^2}$$

$$a = \bar{Y} - \frac{\mu_{11}}{\sigma_X^2}\bar{X}$$

Substituting these in (12.15) we get

$$E\left(\frac{Y}{X}\right) = \bar{Y} - \frac{\mu_{11}}{\sigma_X^2}\bar{X} + \frac{\mu_{11}}{\sigma_X^2}X$$

$$= \bar{Y} + \frac{\mu_{11}}{\sigma_X^2}(X - \bar{X})$$

$$E\left(\frac{Y}{X}\right) = \bar{Y} + r\frac{\sigma_Y}{\sigma_X}(X - \bar{X})$$

Similarly the regression curve of X on Y is

$$E\left(\frac{X}{Y}\right) = \bar{X} + r\frac{\sigma_X}{\sigma_Y}(Y - \bar{Y})$$

Worked Out Examples

EXAMPLE 12.9

A certain city has gathered data on the number of minor traffic accidents and the number of youth soccer games that occur in a town over a weekend.

X (Soccer games)	20	30	10	12	15	25	34
Y (Minor accidents)	6	9	4	5	7	8	9

(i) Predict the number of minor traffic accidents that will occur on a weekend during which 33 soccer games take place in the city.

(ii) Calculate the standard error of estimate.

Solution: Calculations for correlation coefficient

Soccer games (X)	Minor accidents (Y)	X^2	Y^2	XY
20	6	400	36	120
30	9	900	81	270
10	4	100	16	40
12	5	144	25	60
15	7	225	49	105
25	8	625	64	200
34	9	1156	81	306
146	48	3550	352	1101

$$\bar{X} = \frac{\Sigma X}{n} = \frac{146}{7} = 20.86$$

$$\bar{Y} = \frac{48}{7} = 6.86$$

$$\text{cov}(X,Y) = \frac{1}{n}XY - \bar{X}\bar{Y} = \frac{1}{7}(1101) - (20.86)(6.86)$$
$$= 157.285 - 143.099$$
$$= 14.1854$$

$$\sigma_X^2 = \frac{1}{n}\Sigma X^2 - \bar{X}^2 = \frac{1}{7}(3550) - (20.86)^2$$
$$= 507.1428 - 435.1396$$
$$= 72.0032$$
$$= 8.48$$

$$\sigma_Y^2 = \frac{1}{n}\Sigma Y^2 - \bar{Y}^2 = \frac{1}{7}(352) - (6.86)^2$$
$$= 50.285 - 47.0596$$
$$= 3.2254$$
$$= 1.79$$

$$r(X,Y) = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{14.1854}{(8.48)(1.79)} = \frac{14.1854}{15.2295}$$
$$= 0.9314$$

(i) The regression line of Y on X is

$$Y - \bar{Y} = r\frac{\sigma_Y}{\sigma_X}(X - \bar{X}) \Rightarrow Y - 6.86 = \frac{(0.9314)}{8.48}(1.79)(X - 20.86)$$
$$Y - 6.86 = 0.1966(X - 20.86)$$
$$\bar{Y} = 6.86 + 0.1966X - 4.10135$$
$$Y = 0.1966X + 2.7586$$

When $X = 33$, $Y = 9.24 \simeq 9$

Hence, when 33 soccers gave take place, there may be approximately 9 accidents.

(ii) The standard error of estimate of Y is

$$S_Y = \sigma_Y (1 - r^2)^{\frac{1}{2}}$$
$$= 1.79(1 - 0.9314^2)^{\frac{1}{2}}$$
$$= 1.79(0.1324)$$
$$= 0.6515$$

EXAMPLE 12.10

A tire manufacturing company is interested in removing pollutants from the exhaust at the factory and cost is a concern. The company has collected data from other companies concerning the amount of

money spent on environmental measures and the resulting amount of dangerous pollutants released (as a percentage of total emissions).

Money Spent (Y) (₹ Thousands)	8.4	10.2	16.5	21.7	9.4	8.3	11.5	18.4	16.7	19.3	28.4	4.7	12.3
% of Dangerous pollutants (X)	35.9	31.8	24.7	25.2	36.8	35.8	33.4	25.4	31.4	27.4	15.8	31.5	28.9

(i) Calculate the regression lines.

(ii) Predict the percentage of dangerous pollutants released when ₹20, 000 is spent on control measures.

(iii) Calculate the standard error of estimate.

X	Y	$U = \frac{X - 25.4}{10}$	$V = \frac{Y - 16.7}{10}$	U^2	V^2	UV
35.9	8.4	1.05	–0.83	1.1025	0.6889	–0.8715
31.8	10.2	0.64	–0.65	0.4096	0.4225	–0.416
24.7	16.5	–0.07	–0.02	0.0049	0.0004	0.0014
25.2	21.7	–0.02	0.5	0.0004	0.25	–0.01
36.8	9.4	1.14	–0.73	1.2996	0.5329	–0.8322
35.8	8.3	1.04	–0.84	1.0816	0.7056	–0.8736
33.4	11.5	0.8	–0.52	0.64	0.2704	–0.416
25.4	18.4	0	0.17	0	0.0289	0
31.4	16.7	0.6	0	0.36	0.0	0
27.4	19.3	0.2	0.26	0.04	0.0676	0.052
15.8	28.4	–0.96	1.17	0.9216	1.3689	–1.1232
31.5	4.7	0.61	–1.2	0.3721	1.44	–0.732
28.9	12.3	0.35	–0.44	0.1225	0.1936	–0.154
384.0	185.8	5.38	–3.13	6.3548	5.9697	–5.3751

$$\bar{X} = \frac{\Sigma X}{n} = \frac{384}{13} = 29.53$$

$$\bar{Y} = \frac{\Sigma Y}{n} = \frac{185.8}{13} = 14.29$$

$$\bar{U} = \frac{\Sigma U}{n} = \frac{5.38}{13} = 0.4138$$

$$\bar{V} = \frac{\Sigma V}{n} = -0.2047$$

$$\sigma_U^2 = \frac{1}{n}\Sigma U^2 - \bar{U}^2 = \frac{1}{13}(6.3548) - (0.4138)^2 = 0.3176$$

$$\sigma_V^2 = \frac{1}{n}\Sigma V^2 - \bar{V}^2 = \frac{1}{13}(5.9697) - (0.2047)^2 = 0.4173$$

$$\text{cov}(U,V) = \frac{1}{n}\Sigma UV - \bar{U}\bar{V} = \frac{1}{13}(-5.3751) + (0.4138)(0.2047)$$

$$= -0.3287$$

$$r(U,V) = \frac{\text{cov}(U,V)}{\sqrt{\sigma_U^2 \sigma_V^2}} = \frac{-0.3287}{\sqrt{(0.3176)(0.4173)}} = -0.9028$$

Since $r(U, V) = r(X, Y)$. We have $r(X, Y) = -0.9028$

(i) The regression line of Y on X is

$$Y - \bar{Y} = r\frac{\sigma_Y}{\sigma_X}(X - \bar{X})$$

$$Y - 14.29 = (-0.9028)\left(\frac{6.4598}{5.6356}\right)(X - 29.53)$$

$$Y - 14.29 = -1.0348(X - 29.53)$$

$$Y = -1.0348X + 44.848$$

$$\sigma_X^2 = h^2\sigma_U^2$$

$$= 100(0.3176)$$

$$= 31.76$$

$$\sigma_X = 5.6356$$

$$\sigma_Y^2 = k^2\sigma_V^2$$

$$= 100(0.4173)$$

$$= 41.73$$

$$\sigma_Y = 6.4598$$

The regression line of X on Y is

$$X - \bar{X} = r\frac{\sigma_X}{\sigma_Y}(Y - \bar{Y})$$

$$X - 29.53 = (-0.9028)\left(\frac{5.6356}{6.4598}\right)(Y - 14)$$

$$= -0.7876(Y - 14.29)$$

$$= -0.7876Y + 11.2549$$

$$X = -0.7876Y + 40.784$$

(ii) When $Y = ₹20, 000$, $X = ?$

$$X = -0.7876(20, 000) + 40.784$$

$$= -15.7112$$

Hence 15.7112 % of pollutants are released when ₹20, 000 is spent on control measures.

(iii) The standard estimate of Y is given by

$$S_Y = \sigma_Y (1 - r^2)^{\frac{1}{2}}$$
$$= 6.4598[1 - (0.9028)^2]^{\frac{1}{2}}$$
$$= 6.4598(0.43006)$$
$$= 2.7781$$

The standard estimate of X is given by

$$S_X = \sigma_X (1 - r^2)^{\frac{1}{2}}$$
$$S_X = 5.6356[1 - (0.9028)^2]^{\frac{1}{2}}$$
$$= 5.6356(0.43006)$$
$$= 2.4236$$

EXAMPLE 12.11

In a partially destroyed laboratory, record of an analysis of correlation data, the following results are only legible. $\sigma_X^2 = 9$. Regression equations are $8X - 10Y + 66 = 0$ and $40X - 18Y = 214$.

(i) Find mean value of X and Y.

(ii) Find the correlation coefficient between X and Y.

(iii) Find the standard deviation of Y.

Solution:

(i) Since both the regression lines pass through $(\bar{X}, \bar{Y})$, we have $8\bar{X} - 10\bar{Y} + 66 = 0, 40\bar{X} - 18\bar{Y} = 214$. Solving these we get $\bar{X} = 13,\ \bar{Y} = 17$.

(ii) Let $8X - 10Y + 66 = 0, 40X - 18Y = 214$ represent regression lines of Y on X and X on Y, respectively. Hence they can be put in the form:

$$Y = \frac{8}{10}X + \frac{66}{10}, \quad X = \frac{18}{40} + \frac{214}{40}.$$

b_{YX} = regression coefficient of Y on $X = \frac{8}{10} = \frac{4}{5}$

b_{XY} = regression coefficient of X on $Y = \frac{18}{40} = \frac{9}{20}$ $\quad \Rightarrow r^2 = b_{YX} \cdot b_{XY} = \left(\frac{4}{5}\right)\left(\frac{9}{20}\right) = \frac{+3}{5}$

Since both b_{XY}, b_{YX} are positive, $r = \frac{+3}{5}$

(iii) $b_{YX} = r\frac{\sigma_Y}{\sigma_X} \quad \Rightarrow \frac{4}{5} = \frac{3}{5}\frac{\sigma_Y}{3} \quad \Rightarrow \sigma_Y = 4$

EXAMPLE 12.12

Given $f(x, y) = xe^{-x(y+1)}$, $X \geq 0$, $Y \leq 0$

Find the regression curve of Y on X.

Solution: Regression curve of Y on X is given by

$$E(Y/X) = \bar{Y} + r\frac{\sigma_Y}{\sigma_X}(X - \bar{X}) = Y$$

$$E(Y/X) = Y$$

Where

$$E(Y/X) = \int_{-\infty}^{\infty} y\, f(y/x)\, dy = \int_{-\infty}^{\infty} y \frac{f(x,y)}{f_X(x)} dy$$

The marginal probability density of X is

$$f_X(x) = \int_0^{\infty} f(x,y)\, dy = \int_0^{\infty} x\, e^{-x(y+1)}$$

$$f_X(x) = x\, e^{-x} \int_0^{\infty} e^{-xy} dy = x\, e^{-x} \left| \frac{e^{-xy}}{-x} \right|_0^{\infty}$$

$$= e^{-x}, x \geq 0$$

$$f\frac{y}{x} = \frac{f(x,y)}{f_X(x)} = \frac{x\, e^{-x(y+1)}}{e^{-x}} = xe^{-xy}$$

Hence the regression curve of Y on X is

$$y = \int_0^{\infty} y\, f\left(\frac{y}{x}\right) dy = \int_0^{\infty} y\, x\, e^{-xy}\, dy$$

$$= x \int_0^{\infty} y\, e^{-xy}\, dy = x\left[\frac{y\, e^{-xy}}{-x} \Big|_0^{\infty} + \frac{1}{x} \int_0^{\infty} e^{-xy} dy \right]$$

$$y = \frac{e^{-xy}}{-x} \Big|_0^{\infty} = \frac{1}{x}$$

$\therefore xy = 1$ which is a rectangular hyperbola and therefore we can observe that the regression curve is not linear.

EXAMPLE 12.13

The joint density of X and Y is given by

$$f(x,y) = \begin{cases} x + y, & 0 < x < 1,\ 0 < y < 1 \\ 0, & \text{otherwise} \end{cases}$$

(i) Find the correlation coefficient between X and Y.

(ii) The regression curve of Y on X and X on Y.

Solution: The marginal pdf's of X and Y are given by

$$g(x) = \int_{y=0}^{1} f(x,y)\,dy = \int_{0}^{1}(x+y)\,dy = x\,|y|_0^1 + \left|\frac{y^2}{2}\right|_0^1 = x + \frac{1}{2}$$

$$g(x) = x + \frac{1}{2},\ 0 < x < 1$$

$$h(y) = \int_{x=0}^{1} f(x,y)\,dx = \int_{0}^{1}(x+y)\,dx = \left|\frac{x^2}{2}\right|_0^1 + y\,|x|_0^1 = \frac{1}{2} + y$$

$$h(y) = y + \frac{1}{2},\ \ 0 < y < 1$$

The conditional distributions of Y on X and X on Y are given by,

$$f\left(\frac{y}{x}\right) = \frac{f(x,y)}{g(x)} = \frac{x+y}{x+\frac{1}{2}} = \frac{2(x+y)}{2x+1}$$

$$f\left(\frac{x}{y}\right) = \frac{f(x,y)}{h(y)} = \frac{x+y}{y+\frac{1}{2}} = \frac{2(x+y)}{2y+1}$$

The conditional expectation of Y on X and X on Y are given by

$$E\left(\frac{Y}{X}\right) = \int_{y=0}^{1} f\left(\frac{y}{x}\right) dy = \int_{0}^{1} y \cdot \frac{2(x+y)}{2x+1}\,dy = \frac{2}{2x+1}\int_{0}^{1}(xy+y^2)\,dy$$

$$= \frac{2}{2x+1}\int_{0}^{1}\left[x\frac{y^2}{2} + \frac{y^3}{3}\right]_0^1 = \frac{2}{2x+1}\left(\frac{x}{2}+\frac{1}{3}\right) = \frac{(3x+2)}{3(2x+1)}$$

$$E\left(\frac{X}{4}\right) = \int_{x=0}^{1} xf\left(\frac{x}{y}\right)dx = \int_{x=0}^{1} x \cdot \frac{2(x+y)}{2y+1}\,dx$$

$$= \frac{2}{2y+1}\int_{0}^{1}(x^2+xy)\,dx = \frac{2}{2y+1}\left[\left|\frac{x^3}{3}\right|_0^1 + y\left|\frac{x^2}{2}\right|_0^1\right] = \frac{2}{2y+1}\left(\frac{1}{3}+\frac{y}{2}\right)$$

$$= \frac{3y+2}{3(2y+1)}$$

The regression curves for means are

$$y = E\left(\frac{Y}{X}\right) = \frac{3y+2}{3(2y+1)} \text{ and } x = E\left(\frac{X}{Y}\right) = \frac{3y+2}{3(2y+1)}$$

(iii) The marginal densities of X and Y are

$$E(X) = \int_{x=0}^{1} x\, g(x)dx = \int_0^1 x\left(x+\frac{1}{2}\right)dx = \int_0^1\left(x^2+\frac{x}{2}\right)dx$$

$$= \frac{x^3}{3}\Big|_0^1 + \frac{x^2}{4}\Big|_0^1 = \frac{1}{3}+\frac{1}{4} = \frac{4+3}{12} = \frac{7}{12}$$

$$E(X^2) = \int_{x=0}^{1} x^2 g(x)dx = \int_0^1 x^2\left(x+\frac{1}{2}\right)dx = \int_0^1\left(x^3+\frac{x^2}{2}\right)dx$$

$$= \left|\frac{x^4}{4}+\frac{x^3}{6}\right|_0^1 = \frac{1}{4}+\frac{1}{6} = \frac{3+2}{12} = \frac{5}{12}$$

$$V(X) = E(X^2) - [E(X)]^2 = \frac{5}{12} - \left(\frac{7}{12}\right)^2 = 0.0763$$

Similarly we can get

$$E(Y) = \int_{y=0}^{1} y\, h(y)dy = \int_0^1 y\left(y+\frac{1}{2}\right)dy = \int_0^1\left(y^2+\frac{y}{2}\right)dy$$

$$= \left|\frac{y^3}{3}+\frac{y^2}{4}\right|_0^1 = \frac{1}{3}+\frac{1}{4} = \frac{7}{12}$$

$$E(Y^2) = \int_{y=0}^{1} y^2 h(y)\, dy = \int_0^1 y^2\left(y+\frac{1}{2}\right)dy = \int_0^1\left(y^3+\frac{y^2}{2}\right)dy$$

$$= \left|\frac{y^4}{4}+\frac{y^3}{6}\right|_0^1 = \frac{1}{4}+\frac{1}{6} = \frac{5}{12}$$

$$\therefore V(Y) = E(Y^2) - [E(Y)]^2 = \frac{5}{12} - \left(\frac{7}{12}\right)^2 = 0.0763$$

In addition,

$$E(XY) = \int_0^1\int_0^1 xyf(x,y)dxdy = \int_0^1\int_0^1 xy(x+y)dxdy$$

$$= \int_0^1\left[\int_0^1 (x^2y+xy^2)dx\right]dy = \int_0^1\left[y\left|\frac{x^3}{3}\right|_0^1 + y^2\left|\frac{x^2}{2}\right|_0^1\right]dy$$

$$= \int_0^1\left(\frac{y}{3}+\frac{y^2}{2}\right)dy = \left|\frac{y^2}{6}+\frac{y^3}{6}\right|_0^1 = \frac{1}{6}+\frac{1}{6} = \frac{1}{3}$$

$$\text{cov}(X,Y) = E(XY) - E(X)E(Y)$$

$$= \frac{1}{3} - \left(\frac{7}{12}\right)\left(\frac{7}{12}\right) = -0.00694$$

$$r(X,Y) = \frac{\text{cov}(X,Y)}{\sqrt{\sigma_X^2\sigma_Y^2}} = \frac{-0.00694}{\sqrt{(0.0763)(0.0763)}} = 0.0909$$

Work Book Exercises

1. The following table gives the advertisement expenditure in lakhs of rupees against the sales in crores of rupees.

Sales (crores ₹)	14	16	18	20	24	30	32
Advertisement expenditure (₹ lakhs)	52	62	65	70	76	80	78

(i) Estimate the sale for advertising expenditure of ₹100 lakhs.

(ii) Estimate the advertisement expenditure for sales of ₹47 crores.

[**Ans.:** (i) Advertisement expense = ₹102.325 lakhs
(ii) Sale = ₹41.65 crores]

2. From the following data calculate:

(i) The regression coefficients A: $b_{yx} = 7.42$, $b_{xy} = 0.0742$

(ii) Two regression equations A: $y = 7.42x + 6.192$
$x = 0.0742y + 0.6192$

Y \ X	1	2	3	4
10	3	2	–	–
20	2	3	1	–
30	–	1	1	3
40	–	–	3	1

[**Ans.:** (i) $r = 0.742$]

3. The regression lines for two random variables are $3X + 2Y = 26$ and $6X + Y = 31$.

(i) Find $\bar{X}$, $\bar{Y}$.

(ii) Find r correlation coefficient.

(iii) If $\sigma_X^2 = 25$, find Standard deviation of Y from the above data.

[**Ans.:** (i) $\bar{X} = 4$, $\bar{Y} = 7$, (ii) $r = -0.5$, (iii) $\sigma_7 = 15$]

4. For the following set of data:

X	13	16	14	11	17	9	13	17	18	12
Y	6.2	8.6	7.2	4.5	9.0	3.5	6.5	9.3	9.5	5.7

(i) Predict Y for $X = 10, 15, 20$.

(ii) Obtain the standard error of estimate.

[**Ans.:** (i) $Y = 0.7051X - 2.8714$, (ii) $Y = 4.1796, 7.7051, 11.2306$]

5. For the following data, find the regression equations of Y on X and X on Y. In addition, estimate Y when $X = 5.9, 7.2$.

X	2	3	4	5	6
Y	7	9	10	14	15

6. The following results were obtained in the analysis of data on yield of dry bark in ounces (Y) and age in years (X) of 200 cinchona plants.

	X	Y
Average	9.2	16.5
SD	2.1	4.2

Correlation coefficient = 0.84.

(i) Obtain two lines of regression.

(ii) Estimate the dry bark yield of a plant of age 8 years.

7. Given $f(x,y) = \begin{cases} \frac{1}{a^2}, & 0 < x < a, \quad 0 < y < a \\ 0, & \text{otherwise} \end{cases}$

(i) Obtain the regression curve of Y on X.

(ii) Obtain the regression curve of X on Y.

(iii) Obtain the correlation coefficient between X and Y.

8. Suppose $f(x,y) = \begin{cases} \frac{1}{8}(6-x-y), & 0 < x < 2 \\ & 2 < y < 4 \\ 0, & \text{otherwise} \end{cases}$

(i) Obtain regression equation of Y on X and X on Y.

(ii) Obtain r.

(iii) Are X, Y independent?

9. The following table shows the ages (X) and blood pressure (Y) of 8 persons. Obtain the regression equation of Y on X and find the expected blood pressure of a person who is 49 years old.

X	52	63	45	36	72	65	47	25
Y	62	53	51	25	79	43	60	33

[**Ans.:** $Y = 11.87 + 0.768X$
$Y_{49} = 49.502$]

10. In a correlation study, the following values are obtained:

	X	Y
Mean	65	67
SD	2.5	3.5

Coefficient of correlation = 0.8.
Find the two regression equations that are associated with the above data.

11. Two lines of regression are given as $6X + 10Y - 119 = 0$ and $-30X + 45Y + 180 = 0$. The variance of Y is 4.
    (i) Find mean values of X and Y.
    (ii) Find coefficient of correlation between X and Y.
    (iii) The variance of X.

[**Ans.:** (i) $\bar{X} = 12.55 \bar{Y} = 4.37$ (ii) $r = -0.949$ (iii) $\sigma_X^2 = 9.986$]

12. Calculate the correlation coefficient from the following results. $X = 10$, $\Sigma X = 350$, $\Sigma Y = 310$, $\Sigma(X - 35)^2 = 162$, $\Sigma(Y - 31)^2 = 222$, $\Sigma(X - 35)(Y - 31) = 92$. Find the regression line of Y on X.

[**Ans.:** $r = 0.48$, $Y = 41.12 + 0.574$]

DEFINITIONS AT A GLANCE

Regression: A statistical technique to predict one variable from another variable.

Lines of Regression: In a bivariate distribution, if the variables are related such that the points in the scatter diagram will cluster round some line, it is called line of regression.

Regression Coefficient: The slopes of the lines of regression of Y on X and X on Y are called regression coefficients.

Standard Error of Estimate: This measures the variability or scatter of the observed values around the regression line. This is called residual variance.

Regression curve: The conditional mean for a continuous distribution is called regression function of Y on X and its graph is called regression curve of Y on X.

FORMULAE AT A GLANCE

- The regression line of Y on X is

$$Y - \bar{y} = r\frac{\sigma_Y}{\sigma_X}(X - \bar{x})$$

$$= \frac{\mu_{11}}{\sigma_X^2}(X - \bar{x})$$

- The regression line of X on Y is

$$X - \bar{x} = r\frac{\sigma_X}{\sigma_Y}(Y - \bar{y})$$

$$= \frac{\mu_{11}}{\sigma_{Y^2}}(Y - \bar{y})$$

- The regression coefficient of Y on X is

$$b_{YX} = \frac{\mu_{11}}{\sigma_{X^2}} = r\frac{\sigma_Y}{\sigma_X}$$

- The regression coefficient of X on Y is

$$b_{YX} = \frac{\mu_{11}}{\sigma_{y2}} = r\frac{\sigma_X}{\sigma_Y}$$

- If θ is the angle between the two lines of regression then, $\theta = \tan^{-1}\left[\left(\frac{1-r^2}{r}\right)\frac{\sigma_X\sigma_Y}{\sigma_X^2+\sigma_Y^2}\right]$.
- The standard error of estimate of Y is given by $S_Y = \sigma_Y(1-r^2)^{\frac{1}{2}}$
- The standard error of estimate of X is given by

$$S_X = \sigma_X(1-r^2)^{\frac{1}{2}}$$

- The regression curve of Y on X is given by

$$E(Y/X) = \bar{y} + \frac{r\sigma_Y}{\sigma_X}(X - \bar{x})$$

- The regression curve of X on Y is given by

$$E(X/Y) = \bar{x} + r\frac{\sigma_X}{\sigma_Y}(Y - \bar{y})$$

OBJECTIVE TYPE QUESTIONS

1. The regression line of Y on X is given by

(a) $X - \bar{x} = r\frac{\sigma_X}{\sigma_Y}(Y - \bar{y})$ (b) $Y - \bar{y} = r\frac{\sigma_X}{\sigma_Y}(X - \bar{x})$

(c) $Y - \bar{y} = r\frac{\sigma_Y}{\sigma_X}(X - \bar{x})$ (d) none

2. The correlation coefficient between X and Y is given by $r =$ _____.

(a) $\frac{\text{cov}(X,Y)}{\sigma_X}$ (b) $\frac{\text{cov}(X,Y)}{\sigma_Y}$

(c) $\frac{\text{cov}(X,Y)}{\sigma_X\sigma_Y}$ (d) none

3. When $r = 1$, the regression line of X on Y becomes

   (a) $\frac{X-\bar{x}}{\sigma_Y} = \frac{Y-\bar{y}}{\sigma_X}$ (b) $\frac{X-\bar{x}}{\sigma_X} = \pm\frac{Y-\bar{y}}{\sigma_Y}$

   (c) $(X-\bar{x})(Y-\bar{y}) = \sigma_X\sigma_Y$ (d) none

4. The regression coefficient of Y on X is given by b_{YX}

   (a) $\frac{r\sigma_X}{\sigma_Y}$ (b) $r\frac{\sigma_X^2}{\sigma_Y^2}$

   (c) $r\frac{\sigma_Y}{\sigma_X}$ (d) none

5. If $U = \frac{X-a}{h}$, $V = \frac{Y-b}{k}$, then $b_{YX} =$ _____.

   (a) b_{VU} (b) $\frac{k}{h}b_{UV}$

   (c) $\frac{h}{k}b_{UV}$ (d) none

6. Correlation coefficient is the _____ between the regression coefficients.

   (a) geometric mean (b) arithmetic mean

   (c) harmonic mean (d) none

7. If $r = 0.6$, $\sigma_X = 1.2$, $\sigma_Y = 1.8$, $\bar{x} = 8, \bar{y} = 10$ then regression line of Y on X is

   (a) $X = 0.9Y + 2.8$ (b) $Y = 0.9X + 2.8$

   (c) $Y = X + 2.8$ (d) none

8. If $r = 0$ then the angle between the two lines of regression is

   (a) $\theta = \pi$ (b) $\theta = 0$

   (c) $\theta = \frac{\pi}{2}$ (d) none

9. When the two variables have perfect correlation between them, then the two lines of regression

   (a) are at right angles (b) coincide

   (c) are parallel to each other (d) none

10. The standard error of estimate of X is given by $S_X =$ _____.

    (a) $\sigma_Y(1-r)^{\frac{1}{2}}$ (b) $\sigma_X(1-r)^{\frac{1}{2}}$

    (c) $\sigma_X(1-r^2)^{\frac{1}{2}}$ (d) none

11. The regression curve of Y on X is given by $E\left(\frac{Y}{X}\right) =$ _____.

    (a) $r\frac{\sigma_Y}{\sigma_X}(X-\bar{x})$ (b) $\bar{y} + r\frac{\sigma_Y}{\sigma_X}(X-\bar{x})$

    (c) $\bar{x} + r\frac{\sigma_Y}{\sigma_X}(X-\bar{x})$ (d) none

12. If regression line of Y on X is given $6X - 5Y + 33 = 0$ then regression coefficient of Y on X is

(a) $\frac{6}{5}$ (b) $\frac{5}{6}$

(c) $\frac{33}{5}$ (d) none

13. If the two regression lines of Y on X and X on Y are $2X + 3Y = 26$, $Y + 6X = 30$ then $r =$ _____.

(a) -0.9 (b) -0.1

(c) -0.33 (d) none

14. When the two regression lines coincide, then r is

(a) 1 (b) -1

(c) both (a) and (b) (d) none

ANSWERS

1. (c)	2. (c)	3. (b)	4. (c)	5. (b)	6. (a)	7. (b)	8. (c)
9. (b)	10. (b)	11. (b)	12. (a)	13. (b)	14. (c)		

13 Queuing Theory

Prerequisites

Before you start reading this unit, you should:

- Have some knowledge about distributions like Poisson distribution, exponential distribution, and negative exponential distribution
- Have the idea of queues that are formed at various situations
- Have the idea of solving difference questions
- Know the different series available, their sums to *n* terms, and infinity

Learning Objectives

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

- Identify the situations that generate queuing problems
- Understand various elements of a queuing system and each of its description
- Differentiate between different models of queuing theory and their performance measures
- Able to apply the situations to particular models described in the unit

INTRODUCTION

We generally observe in our day-to-day lives, queues at restaurants to eat grocery stores for payment of bills, railway counters to get tickets, and for service in post offices. This phenomena of waiting in queues is not only for human beings but also for cars and vehicles that stop at signal lights, jobs wait to be processed on a machine.

Hence, there is a need to study queues, as it deals with quantifying the phenomenon of waiting in lines using some measures of performance such as average queue length, average waiting time, and average facility utilization.

Queuing theory has been applied to solve many problems. Some of them are as follows:

- Scheduling distribution of scarce war material
- Scheduling of mechanical transport fleets
- Scheduling of jobs in production control
- Minimization of congestion due to traffic delay at toll booths
- Solution of inventory control problems
- Determining the number of attendants required at a petrol station
- Determining the number of runways at an airport
- Scheduling of patients at a clinic
- Determining the size of a parking lot

13.1 ELEMENTS OF A QUEUING MODEL

There are four basic elements in a queuing system or model. They are as follows:

(i) The way in which the queue is formed
(ii) The way in which the server serves
(iii) The queue discipline
(iv) The queue behaviour

The main components of a queuing system are the customers and the server. Customers are generated from a source on arrival at the server, they can start service immediately or wait in a queue if the server is busy. When a server completes a service, it automatically takes the waiting customer. If queue is empty, the server waits till the next customer arrives.

(i) **The way in which a queue is formed:** It is referred to as 'arrival pattern'. From the view of analyzing queues, the arrival pattern is described by inter-arrival time between successive customers. The service is described by the service time per customer. Generally, inter-arrival time and service time are probabilistic or stochastic.

- *Inter-arrival time:* The time between two successive customers is termed as inter-arrival time.
- *Service time:* The time taken by the server to serve the customer is service time.

(ii) **The way in which the server serves:** It is referred to as 'service pattern'. There may be a single server or more servers to serve the customers. The time taken to serve a customer is a random variable. This time may be different for different customers. As an example, one patient may require certain time to be examined by doctor and the next patient may require lesser or more time than the previous one. This time is independent of the time taken by previous patient or next patient.

- *Queue size:* The number of customers in a queue is its size and this may be finite as in the buffer area between two successive machines or it may be infinite as in mail order facilities.

(iii) **Queue discipline:** The order in which the customers are selected from a queue refers to queue discipline which plays important role in the analysis of queuing models.

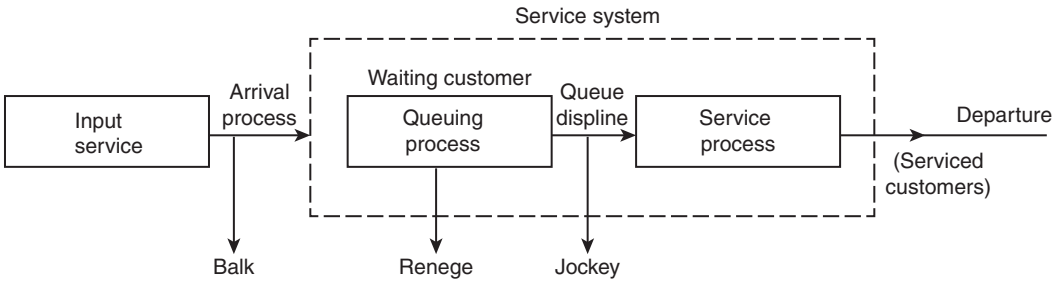
- *FCFS:* This refers to First Come First Served. The customer who comes first is served first. This is observed at railway counters, car garages, etc.
- *LCFS:* This refers to Last Come First Served. The unit which comes last will be served first. This is observed in a godown with all the units stacked. The unit which is lastly placed will be served first.
- *SIRO:* This refers to Service In Random Order. Here, the customers to be served are picked at random for service.
- *SOP:* This refers to Service on Order of Priority. Sometimes though there is a queue waiting to be served, some customers are picked for service according to their order of priority. This is observed in a hospital where emergency cases are served first.

(iv) **Queue behaviour:** The behaviour of customers in a queue refers to queue behaviour. This plays a role in waiting time analysis.

- *Jockey:* Sometimes a customer from the middle of a queue may join a parallel queue with the idea of reducing the waiting time. This process is referred to as jockeying.

- *Balk:* Sometimes a customer may not enter a queue altogether because they anticipate a long waiting or delay. This process of the queue behaviour is referred to as balking.
- *Renege:* Sometimes a customer who has been waiting in a queue for too long time may leave the queue. This process is referred to as reneging.

The queuing system can be represented diagrammatically as follows:



Queuing system

Types of Arrival Patterns

The arrival pattern of customers to the service system is classified into the following two categories:

(i) Static

(ii) Dynamic

These two are further classified based on the nature of arrival rate and the control which can be exercised on the arrival process.

In static arrival patterns, the control depends on arrival rate either random or constant where random arrivals are either at constant rate or varying with time.

In dynamic arrival patterns, the control depends on both service facility and customers. The service facility adjusts its capacity to match changes in the demand intensity by either varying the staffing levels at different timings of service, varying service charges at different things, or allowing entry with appointments.

One of the following probability distributions can be used to approximate arrival time distribution:

(i) Poisson distribution

(ii) Exponential distribution

(iii) Erlang distribution

13.2 DISTRIBUTION OF INTER-ARRIVAL TIME

Consider a Poisson process involving the number, n of arrivals over a time period, t. The probability mass function of Poisson distribution is

$$P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}, x = 0, 1, 2 \ldots$$

If λ is expected number of arrivals per unit time, then expected number of arrivals during a time interval t is (λt).

∴ The probability mass function for this will be

$$P(X = n) = \frac{(\lambda t)^n e^{-\lambda t}}{n!},\ n = 0, 1, 2...$$

The inter-arrival time probability distribution is

$P(X = 0) = \frac{(\lambda t)^0 e^{-\lambda t}}{0!} = e^{-\lambda t}$ which gives the probability of no arrival in the time interval from 0 to t. Let T be the time between successive arrivals which defines a random variable which is continuous as a customer can arrive at any time.

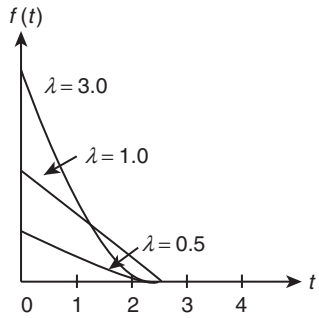
Hence, $P(T > t) = P(X = 0) = e^{-\lambda t}$

Hence there is an arrival in the time interval from 0 to t is

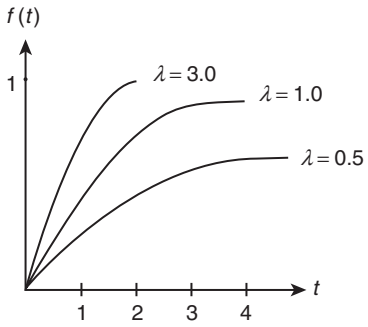
$P(T \le t) = 1 - P(T > t) = 1 - e^{-\lambda t},\ t \ge 0$

The distribution of random variable T is referred to as the exponential distribution whose probability density function is

$$f(t) = \begin{cases} \lambda e^{-\lambda t}, & t \ge 0 \\ 0, & \text{otherwise} \end{cases}$$



Exponential probability density function



Exponential cumulative probability distribution

The mean of the exponential distribution is the expected or average time between arrivals $E(T)$. When these are λ arrivals per unit time, $E(T)=\frac{1}{\lambda}$.

Hence, the Poisson distribution of arrivals with arrival rate λ = negative exponential distribution of inter-arrival times with mean inter-arrival time $\frac{1}{\lambda}$.

The probability density of exponential distribution can also be used to compute the probability that the next customer arrives within time T of the previous arrival, that is, if a customer has already arrived at the service system then the probability of arriving for the next customer can be determined by $F(t) = e^{-\lambda t}$.

13.3 DISTRIBUTION OF SERVICE TIME

The time taken by the server from the beginning of the service to completion of the service for a customer is known as the service time.

If μ is the average service rate, then the expected number of customers served during time interval t will be μt. Thus, if we consider zero time at the start of service, the probability that service is not complete by time t is

$$P(X = 0) = e^{-\mu t}.$$

If T represents the service time which is a random variable, the probability of service completion within time t is given by $P(T \leq t) = 1 - e^{-\mu t}, t \geq 0$.

Hence random service completions lead to a negative exponential distribution of service times.

The mean of this distribution is $E(T) = \frac{1}{\mu}$ which is the average time spent serving a customer.

13.4 QUEUING PROCESS

The queuing process refers to the number of queues and lengths of these queues. There may be single queue or multiple queues.

In certain cases, a service system is unable to accommodate more than the required number of customers at a time. Until some space becomes available, no customers are allowed to enter such situations are referred to as finite source queue. An example can be that of cinema theatres.

In certain other cases, a service system may accommodate any number of customers at a time. These are referred to as infinite source queues. As an example, a departmental store can be taken where in any number of customers are allowed to order and no restriction on the number of orders is placed and hence the size is infinite.

(i) **Queue size:** This is also called line length which gives the total number of customers who are waiting in the queue to be served and not being serviced.

(ii) **Queue length:** The number of customers being served in addition to the line length.

13.5 TRANSIENT STATE AND STEADY STATE

When a system starts its services there may be number of changes. However, it attains stability only after some time. Initially, it is influenced by the initial conditions like number of customers in the queue and elapsed time. This period of transition is called transient state.

However, after sufficient time, the system becomes independent of the initial conditions and of the elapsed time. This state is called steady state.

In queuing theory models, we assume that system has entered a steady state.

13.6 SOME NOTATIONS

The following notations are used in the analysis of queuing system:

n = Number of customers in the system (waiting and being served)

P_n = Probability of n customers in the system

λ = Average number of arrivals per unit time in the queuing system

μ = Average number of customers served per unit time

$$\frac{\lambda}{\mu} = \rho - \frac{\text{Average service completion time}\left(\frac{1}{\mu}\right)}{\text{Average interarrival time}\left(\frac{1}{\lambda}\right)}$$

$\frac{\lambda}{\mu}$ = The expected fraction of time for which the server is busy

s = Number of servers

N = Maximum number of customers allowed in the system

L_s = Expected or average number of customers in the systems both waiting and being served L_s

L_q = Expected or average number of customers in the queue or queue length L_q

W_s = Expected or average waiting time in the system (waiting and being served) W_s

W_q = Expected or average waiting time in the queue W_q

P_w = Probability that an arriving customer has to wait P_w

Worked Out Examples

EXAMPLE 13.1

In a bank operation, the arrival rate is 2 customers per minute. Determine the following:

(i) The average number of arrivals during 5 minutes

(ii) The probability that there are no arrivals in the next 0.5 minutes

(iii) The probability that at least one arrival will occur during the next 0.5 minutes

(iv) The probability that the time between two successive arrivals is at least 3 minutes

Solution:

(i) The arrival rate of customers = $\frac{2}{\text{minutes}}$

During 5 minutes, the average number of arrivals

$$\lambda = 5 \times 2 = 10$$

(ii) Probability that there are no arrivals in the next 0.5 minutes

$$= P(X = 0) = e^{-\lambda t},\ t = 0.5 \text{ minutes}$$

$$P_o(0.5) = P(X = 0) = e^{-10(0.5)} = e^{-5}$$

(iii) Probability that atleast one arrival will occur

$$P(X \geq 1) = 1 - P(X < 1)$$
$$= 1 - P(X = 0)$$
$$= 1 - e^{-5}$$

(iv) Probability that the time between two successive arrivals is at least 3 minutes

$$= P(T > 3),\ t = 3$$
$$= e^{-\lambda t} = e^{-2(3)} = e^{-6}$$
$$= 0.00248$$

EXAMPLE 13.2

On an average 6 customers reach a telephone booth every hour to make calls. Determine the probability that exactly 4 customers will reach 30 minutes period, assuming that arrivals follow Poisson distribution.

Solution: Given average arrivals

$$\lambda = \frac{6 \text{ customers}}{\text{hour}}$$

$t = 30$ minutes $= 0.5$ hour

$n = 4$

$\lambda t = 6 \times 0.5 = 3$ customers

Probability of 4 customers arriving in 0.5 hour

$$= \frac{(\lambda t)^n e^{-\lambda t}}{n!} = \frac{3^4 e^{-3}}{4!} = \frac{81(0.0498)}{24}$$

$$= 0.168$$

13.7 PROBABILITY DISTRIBUTIONS IN QUEUING SYSTEM

Now, we shall look at the distributions that usually exist in queuing systems.

The following axioms are followed when the number of arrivals or departures occurs during an interval of time in a queuing system:

Axiom 1: The probability of an arrival or departure during an interval of time $(t, t + \Delta t)$ depends on the length of time interval Δt and not on either the number of events that occur up to time t or the specific value of t.

Axiom 2: The probability of more than one event occurring during the time interval $(t, t + \Delta t)$ is negligible.

Axiom 3: At most one event can occur during a small interval of Δt. The probability of an arrival during the time interval $(t, t + \Delta t)$ is given by $P_1(\Delta t) = \lambda \Delta t + o(\Delta t)$ where λ is a constant independent of the total number of arrivals up to time t.

13.8 PURE BIRTH PROCESS

The arrival process, in which it is assumed that customers arrive at the queuing system and never leave it, is called pure birth process.

Result: If the arrivals are completely random, then the probability distribution of number of arrivals in a fixed time interval follows a Poisson distribution.

$$P_n(t) = \frac{(\lambda t)^n e^{-\lambda t}}{n!}, \; n = 0, 1, 2, \ldots$$

$P_n(t)$ indicates the number of customers in the system at a time t before the start of the service facility. The mean and variance are equal to λt.

This gives the distribution of arrivals in queuing system.

13.9 PURE DEATH PROCESS

The departure process which assumes that no customers join the system while the service is continued for those who are already in the system is called pure death process. Some axioms are to be followed in

the queuing system. At time $t = 0$, let $N \geq 1$ be the number of customers in the system. Let the service be provided to customers at the rate μ customers leave the system at the rate μ after being serviced.

Axiom-1: The probability of departure during time Δt is $\mu\Delta t$.

Axiom-2: The probability of more than one departure during t and $t + \Delta t$ is negligible.

Axiom-3: The number of departures in non-overlapping intervals are statistically independent.

13.10 CLASSIFICATION OF QUEUING MODELS: (SINGLE SERVER QUEUING MODELS)

Model I: {m/m/1}: {∞/FCFS}

The following assumptions are made about the queuing system:

(i) Poisson distribution of arrivals or exponential distribution of inter-arrival time

(ii) Single waiting line with no restriction on length of queue. No balking or reneging infinite capacity

(iii) Queue discipline FCFS

(iv) Single server with exponential distribution of service time

The system is in state n (number of customers) and no arrival and no departure leaving the total to n customers.

P_n = Probability of being in state n

$$= \frac{(\lambda)^n}{\mu} P_o, \text{ where } P_o = \frac{1-\lambda}{\mu}$$

P_w = Probability that an arriving customer has to wait

$$= 1 - p_o = \frac{\lambda}{\mu} = \rho$$

where P_o denotes the probability of the system being empty (no customer)

Results:

(a) Expected number of customers in the system which are customers in the line plus the customer being served.

$$L_s = \sum_{n=0}^{\infty} n p_n = \sum n(1-\rho)\rho^n \qquad 0 < \rho < 1$$

$$= (1-\rho)\sum_{n=0}^{\infty} n\rho^n (1-\rho)\sum n\rho^n$$

$$L_s = \rho(1-\rho)\sum_{n=1}^{\infty} n\rho^{n-1}$$

$$= \rho(1-\rho)\{1 + 2\rho + 3\rho^2 + \ldots\}$$

$$= \rho(1-\rho)(1-\rho)^{-2}$$

$$\therefore L_s = \frac{\rho}{1-\rho} = \frac{\frac{\lambda}{\mu}}{1-\frac{\lambda}{\mu}} = \frac{\lambda}{\mu-\lambda}, \text{ when } \rho = \frac{\lambda}{\mu}$$

(b) Expected number of customers waiting in the queue or queue length is

$$L_q = \sum_{n=1}^{\infty}(n-1)P_n = \sum_{n=1}^{\infty} n\,p_n - \sum_{n=1}^{\infty} P_n$$

$$= \sum_{n=0}^{\infty} n P_n - \left[\sum_{n=0}^{\infty} P_n - P_o\right] = L_s - (1-P_o)$$

$$= \frac{\lambda}{\lambda-\mu} - \frac{\lambda}{\mu} \quad \text{since } 1-P_o = \frac{\lambda}{\mu}$$

$$L_q = \frac{\lambda^2}{\mu(\mu-\lambda)}$$

(c) Expected waiting time for a customer in the queue is

$W_q = \int_0^{\infty} t\left\{\frac{d}{dt}\phi_w(t)\right\}$, where $\phi_w(t)$ is waiting time distribution of each customer.

$$W_q = \int_0^{\infty} t\lambda(1-\rho)\rho^{-\mu(1-\rho)t}dt$$

Integrating by parts, we get

$$W_q = \lambda(1-\rho)\left[t\frac{\rho^{-\mu(1-\rho)t}}{-\mu(1-\rho)}\right]_0^{\infty} + \int_0^{\infty}\frac{\rho^{-\mu(1-\rho)t\,dt}}{\mu(1-\rho)}$$

$$= \frac{\lambda(1-\rho)}{\mu^2(1-\rho)^2}(\rho^{-\mu(1-\rho)t})_0^{\infty} = \frac{\lambda}{\mu^2\left(1-\frac{\lambda}{\mu}\right)} = \frac{\lambda}{\mu(\mu-\lambda)}$$

$$\therefore W_q = \frac{L_q}{\lambda}$$

(d) Expected waiting time for a customer in the system (waiting and service) is

W_s = Expected waiting time in queue + Expected service time

$$\therefore W_q + \frac{1}{\mu} = \frac{\lambda}{\mu(\mu-\lambda)} + \frac{1}{\mu}$$

$$\therefore W_s = \frac{\lambda+\mu-\lambda}{\mu(\mu-\lambda)} = \frac{1}{\mu-\lambda} = \frac{L_s}{\lambda}$$

(e) Probability that number of customers in the system is greater than or equal to *k*.

$$P(n \geq k) = \left(\frac{\lambda}{\mu}\right)^k \quad \text{since } P(n > k) = \left(\frac{\lambda}{\mu}\right)^{k+1}$$

(f) The variance of queue length

$$\text{var}(n) = \sum_{n=1}^{\infty} n^2 P_n - \left(\sum_{n=1}^{\infty} nP_n\right)^2$$

$$= \sum_{n=1}^{\infty} n^2 P_n - (L_S)^2 = \sum_{n=1}^{\infty} n^2 (1-\rho)\rho^n - \left(\frac{\rho}{1-\rho}\right)^2$$

$$= (1-\rho)\left[1 \cdot \rho^2 + 2^2 \cdot \rho^2 + 3^2 \cdot \rho^3 + \ldots\right] - \left(\frac{\rho}{1-\rho}\right)^2$$

$$\text{var}(n) = \frac{\rho}{(1-\rho)^2} = \frac{\lambda\mu}{(\mu-\lambda)^2}$$

(g) Probability that the queue is non-empty

$$P(n>1) = 1 - P_o - P_1$$

$$= 1 - \left(1 - \frac{\lambda}{\mu}\right) - \left(\frac{1-\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)$$

$$= 1 - \left(1 - \frac{\lambda}{\mu}\right)\left(1 + \frac{\lambda}{\mu}\right) = 1 - 1 + \frac{\lambda^2}{\mu^2}$$

$$= \left(\frac{\lambda}{\mu}\right)^2$$

(h) Probability of k or more customers in the system

$$P(n \geq k) = \sum_{n=k}^{\infty} P_k = \sum_{n=k}^{\infty} (1-\rho)\rho^n$$

$$= (1-\rho)\rho^k \sum_{n=k}^{\infty} \rho^{n-k}$$

$$= (1-\rho)\,\rho^k [1 + \rho + \rho^2 + \ldots] = (1-\rho)\,\rho^k (1-\rho)^{-1}$$

$$= \rho^k \rightarrow P(n \geq k) = \left(\frac{\lambda}{\mu}\right)^k, \quad P(n>k) = \left(\frac{\lambda}{\mu}\right)^{k+1}$$

(i) Expected length of non-empty queue

$$L_b = \frac{\text{Expected length of waiting line}}{\text{Prob}(n>1)}$$

$$= \frac{L_q}{P(n>1)} = \frac{\dfrac{\lambda^2}{\mu(\mu-\lambda)}}{\left(\dfrac{\lambda}{\mu}\right)^2}$$

$$L_b = \frac{\mu}{\mu-\lambda}$$

(j) Probability of an arrival during the service time when system contains r customers

$$P(n=r)=\int_0^\infty P_r(t)s(t)dt$$

$$=\int_0^\infty\left[\frac{(\lambda t)^r e^{-\lambda t}}{r!}\right]\mu e^{-\mu t}\,dt$$

$$=\frac{\lambda^r}{r!}\mu\int_0^\infty t^r e^{-\lambda t-\mu t}dt=\frac{\lambda^r\mu}{r!}\int_0^\infty t^r e^{-(\lambda+\mu)t}dt$$

$$=\frac{\lambda^r\mu}{r!}\frac{\Gamma(r+1)}{(\lambda+\mu)^{r+1}}$$

Since $\int_0^\infty e^{-xt}t^n dt=\frac{\Gamma(n+1)}{x^{n+1}}$ and $\Gamma(\lambda+1)=r!$

$$\therefore\ P(n=r)=\left(\frac{\lambda}{\lambda+\mu}\right)^r\left(\frac{\mu}{\lambda+\mu}\right).$$

Worked Out Examples

EXAMPLE 13.3

A repair shop attended by a single mechanic has an average of four customers per hour who bring small appliances for repair. The mechanic inspects them for defects and quite often can fix them right away or otherwise render a diagnosis. This takes him 6 minutes on the average. Arrivals are Poisson and service time has the exponential distribution.

(i) Find the proportion of time during which the shop is empty

(ii) Find the probability of finding at least one customer in the shop

(iii) Find the average number of customers in the system

(iv) The average time including service spent by a customer

Solution: On an average the mechanic has 4 customers/hours

$\therefore\ \lambda=4$ customers/hour

Since the service is provided at the rate of μ

$$\therefore\ \mu=\frac{60}{6}=\frac{10}{\text{hour}}$$

However,

$$\rho=\frac{\lambda}{\mu}=\frac{4}{10}=0.4$$

(i) The probability that the shop is empty = proportion of time during which stop is empty = P_o

$$=1-\rho=0.6$$

(ii) Probability of finding at least one customer in the shop = Probability that there are more than one customers

$$= P(n \geq 1)$$

$$= \left(\frac{\lambda}{\mu}\right)^1 = (0.4)^1 = 0.4$$

(iii) Average number of customers in the system

$$L_s = \frac{\lambda}{\mu - \lambda} = \frac{4}{10-4} = \frac{4}{6} = \frac{2}{3}$$

(iv) Average time including service spent by a customer

$$W_s = \frac{1}{\mu - \lambda} = \frac{1}{6}\text{hour} = 10 \text{ mintues.}$$

EXAMPLE 13.4

In a bank, cheques are cashed at a single teller counter. Customers arrive at a counter in a Poisson manner at an average rate of 30 customers per hour. The teller takes on an average, a minute, and a half to each cheque. The service time has been shown to be exponentially distributed.

(i) Calculate the percentage of time the teller is busy.

(ii) Calculate the average time a customer is expected to wait.

Solution: Average arrival of customers, $\lambda = \dfrac{30\,\text{customers}}{\text{hour}}$

Average service time for a customer, $\mu = 1\dfrac{1}{2}$ minutes

$$\mu = \frac{60}{\frac{3}{2}} = \frac{40}{\text{hour}}$$

(i) Probability of time the teller is busy = probability that an arriving customer has to wait

$$= 1 - p_o = \frac{\lambda}{\mu} = \frac{30}{40} = \frac{3}{4}$$

(ii) Percentage of time the teller is busy $= \dfrac{3}{4} \times 100 = 75\%$ of time

(iii) Average time a customer is expected to wait

$$W_s = \frac{1}{\mu - \lambda} = \frac{1}{40-30} = \frac{1}{10} \text{ hour}$$

$$= 6 \text{ minutes.}$$

EXAMPLE 13.5

Arrivals at telephone booth are considered to be Poisson with an average time of 10 minutes between one arrival and the next. The length of phone call is assumed to be distributed exponentially, with mean 3 minutes.

(i) What is the probability that a person arriving at the booth will have to wait?

(ii) The telephone department will install a second booth when convinced that an arrival would expect waiting for at least 3 minutes for a phone call. By how much should the flow of arrivals increase in order to justify a second booth?

(iii) What is the average length of the queue that forms from time to time?

(iv) What is the probability that it will take him more than 10 minutes altogether to wait for the phone and complete his call?

Solution: Given that the average time between two arrivals

$$\lambda = \frac{1}{10} = 0.10 \frac{\text{person}}{\text{minute}}$$

Service time for each person (i.e.,) $\mu = \frac{1}{3} = 0.33 \frac{\text{person}}{\text{minute}}$

(i) Probability that a person has to wait at the booth

$$P_w = 1 - p_o = \frac{\lambda}{\mu} = \frac{0.1}{033} = 0.3$$

(ii) The second booth can be installed only if the arrival rate is more than the waiting time.

Let λ^1 be the increased arrival rate. Then the expected waiting time in the queue will be

$$W_q = \frac{\lambda^1}{\mu(\mu - \lambda^1)} \Rightarrow 3 = \frac{\lambda^1}{0.33(0.33 - \lambda^1)}$$

$$0.99(0.33 - \lambda^1) = \lambda^1$$

$$0.3267 = \lambda^1(1 + 0.99)$$

$$\lambda^1 = 0.164$$

Hence, the increase in the arrival rate $= \lambda^1 - \lambda$

$$= 0.164 - 0.10$$

$$= 0.06 \frac{\text{arrivals}}{\text{minute}}$$

Here, given that waiting time, $W_q = 3$ minutes

(iii) Average length of non-empty queue $= L_b$

$$L_b = \frac{\mu}{\mu - \lambda} = \frac{0.33}{0.33 - 0.1} = \frac{0.33}{0.23}$$

$L_b = 2$ customers approximately

(iv) Probability of waiting for more than 10 minutes is

$$P(t \geq 10) = \int_{10}^{\infty} \frac{\lambda}{\mu}(\mu - \lambda)\, e^{-(\mu - \lambda)t} dt$$

$$= \int_{10}^{\infty} \frac{0.1}{0.33}(0.33 - 0.1) e^{-(0.33 - 0.1)t} dt$$

$$P(t \geq 10) = 0.069 \frac{e^{-0.23t}}{-0.23}\Big|_{10}^{\infty} = 0.03$$

This shows that 3% of the arrivals on an average will have to wait for 10 minutes or more before they can use the phone call.

EXAMPLE 13.6

Trucks arrive at a factory for collecting finished goods for transportation to distant markets. As and when they come they are required to join a waiting line and are served on FCFS basis. Trucks arrive at the rate of 10 per hour whereas the loading rate is 15 per hour.

It is also given that arrivals are Poisson and loading is exponentially distributed. Transporters have complained that their trucks have to wait for nearly 12 hours at the plant.

(i) Examine whether the complaint is justified

(ii) In addition, determine probability that the loaders are idle in the above problem.

Solution: Trucks arrive on an average, $\lambda = \frac{10}{\text{hour}}$

The loading rate of trucks, $\mu = \frac{15}{\text{hour}}$

(i) Average waiting time for the trucks

$$W_q = \frac{\lambda}{\mu(\mu - \lambda)} = \frac{10}{15(15-10)} = \frac{10}{75} \text{ hours}$$

$W_q = 8$ minutes approximately

The complaint is not justified as the waiting time for each is just 8 minutes approximately.

(ii) Probability that the loaders are idle, that is, probability that system is idle (Idle time) $= P_o = 1 - \rho$

$$= 1 - \frac{\lambda}{\mu} = 1 - \frac{10}{15} = \frac{5}{15}$$

$$= 33.33\%$$

The loaders are idle for 33.33% of times.

EXAMPLE 13.7

Workers come to tool store room to receive special tools required by them for accomplishing a particular project assigned to them. The average time between two arrivals is 60 seconds and the arrivals are assumed to be in Poisson distribution. The average service time of the tool room attendant is 40 seconds. Determine:

(i) The average queue length

(ii) Average length of non-empty queues

(iii) Average number of workers in system including the worker being attended

(iv) Mean waiting time of an arrival

(v) Average waiting time of an arrival of a worker who waits

(vi) The type of policy to be established. In other words, determine whether to go for an additional tool store room attendant which will minimize the combined cost of attendant's idle time and the cost of workers' waiting time. Assume that the charges of a skilled worker is ₹4 per hour and that of tool store room attendant is ₹0.75 per hour.

Solution: Given that average time between two arrivals

$$\lambda = \frac{1}{60} \text{ per second} = 1 \text{ per minute}$$

The average service time = 40 seconds or

$$\mu = \frac{1}{40} \text{ per second} = 1.5 \text{ per minute}$$

(i) Average queue length, $L_q = \frac{\lambda}{\mu} \cdot \frac{\lambda}{\mu - \lambda}$

$$= \frac{1}{1.5} = \frac{1}{1.5 - 1}$$

$$= \frac{1}{0.75} = \frac{4}{3} \text{ workers}$$

(ii) Average length of non-empty queues is

$$L_b = \frac{\mu}{\mu - \lambda} = \frac{1.5}{1.5 - 1} = 3 \text{ workers}$$

(iii) Average number of workers in the system,

$$L_s = \frac{\lambda}{\mu - \lambda} = \frac{1}{1.5 - 1} = \frac{1}{0.5} = 2 \text{ workers}$$

(iv) Mean waiting time for an arrival

$$W_q = \frac{\lambda}{\mu} \cdot \frac{1}{\mu - \lambda} = \frac{1}{1.5} \cdot \frac{1}{1.5 - 1} = \frac{4}{3} \text{ minutes}$$

(v) Average waiting time of an arrival who waits (waiting + service)

$$W_s = \frac{1}{\mu - \lambda} = \frac{1}{1.5 - 1} = \frac{1}{0.5} = 2 \text{ minutes}$$

(vi) Probability that the tool room attendant remains idle

$$P_o = 1 - \rho = 1 - \frac{\lambda}{\mu} = 1 - \frac{1}{1.5} = \frac{0.5}{1.5} = \frac{1}{3}$$

Idle time cost of one attendant $= \frac{1}{3} \times 8 \times 0.75 = \frac{₹2}{\text{day}}$

Waiting time cost of workers = W_q × no. of workers arriving per day × cost of worker

$$= \left(\frac{4}{3} \times \frac{1}{16}\right)(8 \times 60) \times ₹4$$

$$= ₹\frac{128}{3} = \frac{₹42.67}{\text{day}}.$$

EXAMPLE 13.8

A fertilizer's company distributes its products by trucks loaded at its only loading station. Both the company trucks and contractor's trucks are used for this purpose. It was found out that on an average every 5 minutes one truck arrived and average loading time was 3 minutes. 40% of the trucks belong to the contractors. Determine:

(i) The probability that a truck has to wait

(ii) The waiting time of the truck that waits

(iii) The expected waiting time of customers' trucks per day.

Solution: Average arrival of trucks = one for 5 minutes

$$\therefore \lambda = \frac{60}{5} = 12 \frac{\text{trucks}}{\text{hour}}$$

Average service time for each truck, $\mu = \frac{60}{3} = 20 \frac{\text{trucks}}{\text{hour}}$

(i) Probability that a truck has to wait:

$$P_w = 1 - P_o = \frac{\lambda}{\mu} = \frac{12}{20} = 0.6$$

(ii) Waiting time of truck that waits (waiting + service):

$$W_s = \frac{1}{\mu - \lambda} = \frac{1}{20 - 12} = \frac{1}{8} \text{ hours} = 7.5 \text{ minutes}$$

(iii) Expected waiting time of customers trucks per day

= Total waiting time per day

= (Number of trucks per day) × (percent contractors trucks) × Expected waiting time for a truck

$$= (12 \times 24) \times \frac{40}{100} \times \frac{\lambda}{\mu(\mu - \lambda)}$$

$$= 12 \times 24 \times \frac{4}{10} \times \frac{12}{20(20 - 12)} = 8.64 \text{ hours per day}$$

EXAMPLE 13.9

Customers arrive at the first class ticket counter of a theatre at the rate of 12 per hour. There is one clerk serving the customers at the rate of 30 per hour.

(i) What is the probability that there is no customer in the counter, that is, system is idle?

(ii) What is the probability that there are more than 2 customers in the counter?

(iii) What is the probability that there is no customer waiting to be served?

(iv) What is the probability that a customer is being served and nobody is waiting?

Solution: Given that, arrival of customers, = 12 per hour $\lambda = 12$ per hour

Service rate of the customers, $\mu = \frac{30}{\text{hour}}$

(i) Probability that there is no customer in the queue

$$= P_o = 1 - \rho = 1 - \frac{\lambda}{\mu} = 1 - \frac{12}{30} = \frac{8}{30} = 0.6$$

(ii) Probability that there are more than two customers in the counter

$$= P_3 + P_4 + P_5 + \ldots$$

$$= 1 - (P_0 + P_1 + P_2)$$

$$= 1 - \left[P_0 \left(1 + \frac{\lambda}{\mu} + \frac{\lambda^2}{\mu^2} \right) \right]$$

$$= 1 - \left[0.6 \left(1 + \frac{12}{30} + \frac{144}{900} \right) \right]$$

$$= 1 - [0.6(1 + 0.4 + 0.16)]$$

$$= 1 - 0.936 = 0.064$$

(iii) Probability that there is no customer waiting to be served = Probability that there is at most one customer in the counter being serviced

$$= P_0 + P_1 = 0.6 + 0.6 \left(\frac{12}{30} \right)$$

$$= 0.84$$

(iv) Probability that a customer is being served and nobody is waiting

$$= P_1 = P_o \cdot \frac{\lambda}{\mu} = 0.6 \left(\frac{12}{30} \right)$$

$$= 0.24$$

EXAMPLE 13.10

The tool room company's quality control department is managed by a single clerk who takes an average of 5 minutes in checking parts of each machine coming for inspection. The machines arrive once in every 8 minutes on the average one hour of the machine is valued at ₹15 and the clerk's time is valued at ₹4 per hour. What is the cost of average hourly queuing system associated with the quality control department?

Solution: Average number of arrivals, λ = once in every 8 minutes

$$= \frac{60}{8} = \frac{7.5}{\text{hour}}$$

Average service time for each machines, $\mu = 5$ minutes

$$\mu = \frac{60}{5} = \frac{12}{\text{hour}}$$

Average waiting time for a customer in the system

$$W_s = \frac{1}{\mu - \lambda} = \frac{1}{12 - 7.5} = \frac{2}{9} \text{ hours}$$

Average queuing cost per machine $= ₹\left(15 \times \frac{2}{9}\right)$

$$= ₹\frac{10}{3}$$

Average queuing cost per hour $= ₹\frac{10}{3} \times 7.5$

$= ₹25$

Average cost of a clerk per hour $= ₹4$

$\therefore$ Total cost for quality control department $= ₹25 + ₹4$

$= ₹29$

EXAMPLE 13.11

On the average 96 patients per 24-hour day require the service of an emergency clinic. In addition, on the average, a patient requires 10 minutes of active attention. Assume that facility can handle only one emergency at a time. Suppose that it costs the clinic ₹100 per patient treated to obtain an average servicing time of 10 minutes, and that each minute of decrease in this average time would cost the clinic ₹10 per patient treated. How much would have to be budgeted by the clinic to decrease the average size of the queue from $1\frac{1}{3}$ patients to $\frac{1}{2}$ patient?

Solution: Average arrivals, $\lambda = \frac{96}{24}$

$$= \frac{4 \text{ patients}}{\text{hour}}$$

Average service rate, $\mu = \frac{1}{10} \times 60$

$$= \frac{6 \text{ patients}}{\text{hour}}$$

Average number of patients in the queue = queue length

$$L_q = \frac{\lambda}{\mu} \cdot \frac{\lambda}{\mu - \lambda} = \frac{4}{6} \cdot \frac{4}{6 - 4} = \frac{4}{3}$$

$$= 1\frac{1}{3}$$

This number is to be reduced from $1\frac{1}{3}$ to $\frac{1}{2}$

Let $L_q^1 = \frac{1}{2}$ which can be achieved by increasing the service rate to μ^1

$$L_q^1 = \frac{\lambda}{\mu^1} \cdot \frac{\lambda}{\mu^1 - \lambda} \Rightarrow \frac{1}{2} = \frac{4}{\mu^1} \cdot \frac{4}{\mu^1 - 4}$$

$$\mu^1(\mu^1 - 4) = 32$$

$$\mu_1^2 - 4\mu^1 - 32 = 0 \Rightarrow (\mu^1 - 8)(\mu^1 + 4) = 0$$

$$\mu^1 = 8 \text{ or } \mu^1 = -4$$

$\therefore \mu^1 = 8 \because \mu^1$ cannot be negative.

$\therefore$ Average time required per each patient $= \frac{1}{8}$ hour $= \frac{15}{2}$ minutes

Decrease in the time required to attend a patient

$$= 10 \frac{-15}{2} = \frac{5}{2} \text{ minutes}$$

Hence, budget required for each patient $= ₹100 \times \frac{5}{2} \times 10$

$= ₹125$

Hence to decrease the size of the queue, the budget per patient should be increased from ₹100 to ₹125.

Model II: {(m/m/1): (∞/SIRO)}

- Exponential distribution of interval time or Poisson distribution of arrivals
- Single waiting line
- No restriction on minimum number of customers
- Queue discipline is SIRO

Hence, this model is identical to model I with a difference only in the queue discipline.
Since P_n is independent of any specific queue discipline.
Here also $P_n = (1-\rho)\,\rho^n$, $n = 1, 2, \ldots$
All other results are the same as in model I.

Worked Out Examples

EXAMPLE 13.12

If for a period of 2 hours in the day(8 to 10 am) trains arrive at the yard every 20 minutes but the service time continues to remain 36 minutes. For this period, calculate:

(i) The probability that the yard is empty

(ii) The average number of trains in the system on the assumption that line capacity of the yard is limited to μ trains only

Solution: Arrival rate of trains, $\lambda = \frac{1}{20}$

Service rate of trains, $\mu = \frac{1}{36}$

$\therefore \frac{\lambda}{\mu} = \rho = \frac{36}{20} = 1.8$ which is greater than 1.

(i) Probability that the yard is empty = $P_o = \frac{\rho - 1}{\rho^{N+1} - 1}$

where N denotes the customers already present in the system = 4

$$\therefore P_o = \frac{1.8-1}{1.8^{4+1}-1} = \frac{0.8}{17.89568} = 0.0447$$

(ii) Average number of trains in the system

$$L_s = \sum_{n=0}^{4} nP_n = P_1 + 2P_2 + 3P_3 + 4P_4$$

$$= \frac{(1-\rho)\rho}{1-\rho^{N+1}} + \frac{2(1-\rho)\rho^2}{1-\rho^{N+1}} + \frac{3(1-\rho)\rho^3}{1-\rho^{N+4}} + \frac{4(1-\rho)\rho^4}{1-\rho^{N+1}}$$

$$\therefore L_s = \left(\frac{1-\rho}{1-\rho^{N+1}}\right)[\rho + 2\rho^2 + 3\rho^3 + 4\rho^4]$$

$$= \left(\frac{1-1.8}{1-(1.8)^5}\right)[1.8 + 2(1.8)^2 + 3(1.8)^3 + 4(1.8)^4]$$

$$= 2.9 \simeq 3$$

On an average the system has 3 trains.

Model III: {(m/m/1): (N/FCFS)}

- The capacity of the system is N, finite in number
- Queue discipline is FCFS

This is different from model I. Here, not more than N customers can be accommodated at any time in the system. Thus, any customer who arrives when there are N customers in the system does not enter the system. As an example, an emergency room in a hospital cannot take beyond certain number.

Results:

(a) Expected number of customers in the system

$$L_s = \sum_{n=1}^{N} nP_N = \sum_{n=1}^{N} n\left(\frac{1-\rho}{1-\rho^{N+1}}\right)\rho^n$$

$$= \frac{1-\rho}{1-\rho^{N+1}} \sum_{n=1}^{N} n\rho^n \frac{1-\rho}{1-\rho^{N+1}}\left[\rho + 2\rho^2 + 3\rho^3 + \ldots N\rho^N\right]$$

$$L_s = \begin{cases} \dfrac{\rho}{1-\rho} - \dfrac{(N+1)\rho^{N+1}}{1-\rho^{N+1}}; & \rho \neq 1(\lambda \neq \mu) \\ \dfrac{N}{2}; & \rho = 1(\lambda = \mu) \end{cases}$$

(b) Expected number of customers waiting in the system = Expected queue length

$$L_q = L_s - \frac{\lambda}{\mu} = L_s - \frac{\lambda(1-P_N)}{\mu}$$

(c) Expected waiting time of a customer in the system (waiting + service),

$$W_s = \frac{L_s}{\lambda(1-P_N)}$$

(d) Expected waiting time of a customer in the queue

$$W_q = W_s - \frac{1}{\mu} = \frac{L_q}{\lambda(1-P_N)}$$

(e) Fraction of time the system is full or fraction of potential customers lost is

$$P_N = P_o \rho^N$$

Effective arrival rate, $\lambda_e = \lambda(1 - P_N)$

Effective traffic intensity, $\rho_e = \frac{\lambda_e}{\mu}$

EXAMPLE 13.13

Patients arrive at a clinic according to Poisson distribution at the rate of 30 patients per hour. The waiting room does not accommodate more than 14 patients. Examination time per patient is exponential with mean rate of 20 per hour.

(i) Find effective arrival rate at the clinic.

(ii) What is the probability that an arriving patient will not wait and will he find a vacant set in the room?

(iii) What is the expected waiting time until a patient discharged from the clinic?

Solution: Arrival rate of patients, $\lambda = \frac{30}{60} = \frac{1}{2}$

Average service rate, $\mu = 20$ per hour $= \frac{20}{60} = \frac{1}{3}$

$$\therefore \rho = \frac{\lambda}{\mu} = \frac{3}{2} = 1.5$$

Maximum number of customers that can be accommodated $N = 14$

(i) Effective arrival rate at the clinic, $\lambda_e = \lambda(1-P_N) = \lambda(1-P_o e^N)$

$$P_o = \frac{(\rho-1)\rho^n}{\rho^{N+1}-1} = \frac{(1.5-1)(1.5)^0}{(1.5)^{14+1}-1}$$

$$= \frac{0.5}{11221.74} = 0.0000445$$

(ii) Probability that an arriving patient will not wait

= Probability that queue is empty

$= P_o = 0.0000445$

Hence, there are 0.0004% chances that the patient will not wait and so he will not get a seat.

(iii) Expected waiting until a patient is discharged from the clinic is $W_s = \frac{L_s}{\lambda(1-P_N)}$

$$L_s = \frac{\rho}{1-\rho} - \frac{(N+1)\rho^{N+1}}{1-\rho^{N+1}}$$

$$= \frac{1.5}{1.5-1} \cdot \frac{15(1.5)^{15}}{(1.5)^{15}-1} = 3.00026$$

$$P_N = P_o \rho^N = \frac{0.5}{11221.74} \times (1.5)^{14}$$

$$= 0.33336$$

$$\therefore W_s = \frac{3.00026}{0.5(0.66664)} = \frac{3.00026}{0.33332}$$

$$= 9.0011$$

Expected waiting time is 9 hours until a patient is discharged from the clinic.

EXAMPLE 13.14

In a car wash service facility, cars arrive for service according to Poisson distribution with mean 5 per hour. The time for washing and cleaning each car has exponential distribution with mean 10 minutes per car. The facility cannot handle more than one car at a time and has a total of 5 parking spaces.

(i) Find the effective arrival rate.

(ii) What is probability that an arriving car will get service immediately upon arrival?

(iii) Find the expected number of parking spaces occupied.

Solution: Average arrival rate, $\lambda = 5$ per hour

Average service rate for each car, $\mu = \frac{60}{10}$

$= 6$ per hour

Number of cars that can be accommodated

$$N = 5 \Rightarrow \rho = \frac{5}{6} = 0.333$$

(i) Effective arrival rate, $\lambda_e = \lambda(1 - P_N)$

$$= \lambda(1 - P_o \rho^N)$$

$$P_o = \frac{1-\rho}{1-\rho^{N+1}} = \frac{1-0.33}{1-(0.33)^6} = \frac{0.67}{0.9987} = 0.67$$

$$\therefore \lambda_e = 5(1 - (0.67)\ (0.33)^5)$$

$$= 4.98 \simeq 5$$

(ii) Probability that an arriving car will get service immediately upon arrival $= P_o = 0.67$

(iii) Expected number of parking spaces occupied

$$= L_s = \sum_{n=1}^{N} nP_n = \sum \frac{n(1-\rho)\rho^n}{1-\rho^{N+1}}$$

$$= \frac{(1-\rho)}{1-\rho^{N+1}} \sum_{n=1}^{5} ne^n = \frac{1-0.33}{1-(0.33)^6}[\rho + 2\rho^2 + 3\rho^3 + 4\rho^4 + 5\rho^5]$$

$$= \frac{0.67}{0.9987}\left[0.33 + 2(0.33)^2 + 3(0.33)^3 + 4(0.33)^4 + 5(0.33)^5\right]$$

$$= [0.33 + 0.2178 + 0.10781 + 0.04743 + 0.01956]$$

$$= 0.484$$

Work Book Exercises

Model I & II:

1. A television repairman finds that the time spent on his jobs has an exponential distribution with a mean of 30 minutes. If he repairs sets in order in which they came in and if the arrival of sets follows a Poisson distribution approximately with an average rate of 10 per 8 hour day.
   (i) What is the repairman's expected idle time each day?
   (ii) How many jobs are ahead of the average set just brought in?

   [**Ans.:** (i) 3 hours (ii) 2 TV sets (approximately)]

2. A maintenance service facility has Poisson arrival rates, negative exponential service times, and operates on a FCFS discipline. Break downs occur on an average of three per day with a range of zero to eight. The maintenance crew can service on an average of six machines per day with a range from zero to seven. Find:
   (i) Utilization factor of service facility
   (ii) Mean waiting time in the system
   (iii) Mean number of machines in the system
   (iv) Probability of finding 2 machines in the system.

   [**Ans.:** $\lambda = \frac{3}{\text{day}}, \mu = \frac{6}{\text{day}}$, (i) $\rho = \frac{\lambda}{\mu} = \frac{3}{6}$ or 50% (ii) $W_s = \frac{1}{3}$ day

   (iii) $L_s = 1$ machine (iv) $W_q = 1.6$ day, (v) $P_2 = \left(\frac{\lambda}{\mu}\right)^2\left(1 - \frac{\lambda}{\mu}\right) = 0.125$]

3. Customers arrive at a box office window being manned by a single individual according to a Poisson input process with a mean rate of 30 per hour. The time required to serve a customer has an exponential distribution with a mean of 90 seconds.
   (i) Find the average waiting time of a customer.
   (ii) In addition, determine the average number of customers in the system.
   (iii) Find the average queue length.

   [**Ans.:** $\lambda = \frac{30}{60}, \ \mu = \frac{60}{90}$ (i) $W_s = \frac{4.5 \text{ minutes}}{\text{customers}}$ $L_s = 3$, (iii) $L_q = 0.25$]

4. A person repairing radios finds that time spent on the radio sets has exponential distribution with mean 20 minutes. If the radios are repaired in the order in which they come in, and their arrival is approximately Poisson with an average rate of 15 for 8 hour per day.

(i) What is the repairman's expected idle time each day?

(ii) How many jobs are ahead of the average set just brought in?

(iii) What is expected waiting time for each customer?

[**Ans. :** $\lambda = \frac{1}{32}, \mu = \frac{1}{20}$ (i) $L_s = \frac{5}{3}$,

(ii) No. of hours $= 8\frac{\lambda}{\mu} = 5$ hours (iii) Time = 8 − 5 = 3 hours.]

5. A branch of Punjab National Bank has only one typist. Since the typing work varies in length (number of pages to be typed), the typing rate is randomly distributed approximating a Poisson distribution with mean service rate of 8 letters per hour. The letters arrive at a rate of 5 per hour during the entire 8 hour work day. If the typewriter is valued at ₹1.50 per hour, find:

(i) Equipment utilization

(ii) The per cent time that an arriving letter has to wait

(iii) Average system time

(iv) Average cost due to waiting on the part of typewriter, that is, it remains idle.

[**Ans. :** $\lambda = \frac{5}{\text{hour}}, \mu = \frac{8}{\text{hour}}$, (i) $\rho = \frac{\lambda}{\mu} = 0.625$

(ii) 62.5% (iii) $W_s = \frac{1}{3}$ hour = 20 min (iv) $\cos t = 8\left(1 - \frac{5}{8}\right) \times ₹1.50 = ₹4.50$]

6. The mean rate of arrival of planes at an airport during the peak period is 20/hour as per Poisson distribution. During congestion, the planes are forced to fly over the field in stack awaiting the landing of other planes that had arrived earlier.

(i) How many planes would be flying in the stack during good and in bad weather?

(ii) How long a plane would be in stack and in the process of landing in good and in bad weather?

(iii) How much stack and landing time to allow so that priority to land out of order would have to be requested only once in 20 times?

[**Ans.:** $\lambda = \frac{20 \text{ planes}}{\text{hour}}, \mu = \frac{60 \text{ planes}}{\text{hour}}$, (i) $L_q = \frac{1}{6}$ planes, $W_S = 1.5$ minutes

(ii) $\int_0^t (\mu - \lambda) + e^{-(\mu-\lambda)t} dt = 0.95$ which gives $t = 0.075$ hours = 4.5 minutes.

In bad weather, $L_q = \frac{4}{3}$ planes, $W_s = 6$ minutes, $t = 0.3$ hour = 18 minutes]

7. Barber *A* takes 15 minutes to complete on haircut. Customers arrive in his shop at an average rate of one every 30 minutes. Barber *B* takes 25 minutes to complete one haircut and customers arrive at his shop at an average rate of one every 50 minutes. The arrival processors are Poisson and the service times follow an exponential distribution.

(i) Where would you expect a bigger queue?

(ii) Where would you require more time waiting included to complete a haircut?

[**Ans.:** Barber *A*, $\lambda = \frac{1}{30}$ minute, $\mu = \frac{1}{15}$ minute, $\rho = \frac{1}{2}$ $L_q = \frac{1}{2}, W_S = 30$ minutes.

At Barber shop B, $\lambda = \frac{1}{50}$, $\mu = \frac{1}{25}$, $\rho = \frac{1}{2}$, $L_q = \frac{1}{2}$, $W_s = 50$ mintues]

8. A toll gate is operated on a freeway where cars arrive according to a Poisson distribution with mean frequency of 1.2 cars per minute. The time of completing payment follows an exponential distribution with mean of 20 seconds. Find:
   (i) The idle time of the counter
   (ii) Average number of cars in the system
   (iii) Average no of cars in the queue
   (iv) Average time that a car spends in the system
   (v) Average time that a car spends in the queue
   (vi) The probability that a car spends more than 30 seconds in the system

   [**Ans.:** $\lambda = \frac{1.2 \text{ cars}}{\text{minute}} = 0.02$ per second, $\mu = \frac{1}{20}$ $\rho = \frac{2}{5}$ (i) $P_o = \frac{3}{5}$, (ii) $L_s = \frac{2}{3}$, (iii) $L_q = \frac{4}{15}$,
   (iv) $W_s = \frac{100}{3}$ sec, (v) $W_q = \frac{40}{3}$ seconds, (vi) $\int_{30}^{\infty} (\mu - \lambda) e^{-(\mu-\lambda)t} dt = 0.4065$]

9. Customers arrive at the first class ticket counter of theatre at a rate of 12 per hour. There is one clerk serving the customers at the rate of 30 per hour. The arrivals are Poisson in nature and the service time follows exponential distribution. Find
   (i) Probability that there is no customer at the counter
   (ii) Probability that there is no customer waiting to be served
   (iii) Probability that a customer is being served and nobody is waiting.

10. An *E*-seva Kendra in a small town has only one bill receiving window with cashier handling the cash transaction and giving receipts. He takes an average 5 minutes per customer. The customers come at random with an average of 8 per hour and the arrivals are Poisson in nature. Determine:
   (i) Average queue length
   (ii) Expected idle time of the cashier
   (iii) Expected time a new arrival spends in the system
   (iv) Expected waiting time of a new arrival before his service is started
   (v) Probability that a person has to spend for at least 10 minutes in the system

Model III:

11. A Petrol station has a single pump and space not more than 3 cars (2 waiting, 1 being served). A car arriving when the space is filled to capacity goes elsewhere for petrol. Cars arrive according to Poisson distribution at a mean rate of one every 8 minutes. Their service time has an exponential distribution with a mean of 8 minutes. The owner has the opportunity of renting an adjacent piece of land which would provide space for an additional car to wait. The rent would be ₹2000 per month. The expected net profit from each customer is ₹2 and the station is open 10 hours every day. Would it be profitable to rent the additional space?

12. At a railway station, only one train is handled at a time. The railway yard is sufficient only for two trains to wait, while the other is given signal to leave the station. Trains arrive at the station at an average rate of 6 per hour and the railway station can handle then on an average of 12 per hour. Assuming Poisson arrivals and exponential service distribution:
    (i) Find the steady state probability has for the various number of trains in the system.
    (ii) In addition, find the average waiting time of a new train coming into the yard.

    [**Ans.:** $\lambda = 6, \mu = 12, \rho = 0.5, N = 3, P_o = 0.53, L_s = \sum_{n=1}^{3} nP_n = 0.74, L_q = 0.24, W_q = 0.04$]

13.11 MULTI-SERVER QUEUING MODELS

Model IV: {(m/m/s): (∞FCFS)}

- Inserted of a single server as in model I, there are multiple servers in parallel which are 's' in number.
- Queue discipline is FCFS.
- Arrival rate of customers follows Poisson distribution with an average rate of λ customers per unit of time and service rate distribution exponential with an average of μ customers per unit of time.

(a) If $n < s$, number of customers in the system are less than the number of servers, there will be no queue. Number of servers are not busy and combined service rate will be $\mu_n = n_\mu$, $n < s$.

(b) If $n \geq s$, number of customers in the system is more than or equal to the number of servers, then all servers will be busy and the maximum number of customers in the queue will be $(n - s)$. Then combined service rate will be

$$\mu_n = s_\mu,\ n \geq s.$$

Results:

(i) Expected number of customers waiting in the queue (length of line) is $L_q = \sum_{n=s}^{\infty} (n-s)P_n$

$$= \sum_{n=s}^{\infty} (n-s)\frac{\rho^n}{s^{n-s}s!}P_o$$

$$L_q = \frac{\rho^s P_o}{s!}\sum_{n=s}^{\infty}(n-s)\rho^{n-s}$$

$$= \frac{\rho^s}{s!}P_o \sum_{m=0}^{\infty} m\rho^m \quad \text{by taking } n - s = m$$

$$= \frac{\rho^s}{s!}\rho P_o \sum_{m=0}^{\infty} me^{m-1}$$

$$= \frac{\rho^s}{s!}\rho P_o \frac{d}{d\rho}\sum_{m=1}^{\infty} e^m \quad \text{from the results of difference equations.}$$

$$= \frac{\rho^s}{s!}\rho P_o \frac{1}{(1-\rho)^2} = \frac{\left(\frac{\lambda}{\mu}\right)^s}{s!} P_o \frac{1}{\left(1-\frac{\lambda}{\mu}\right)^2} \cdot \frac{\lambda}{\mu}$$

$$= \left[\frac{1}{s!}\left(\frac{\lambda}{\mu}\right)^s \frac{\lambda s\mu}{(s\mu-\lambda)^2}\right] P_o \quad \text{since } \rho = \frac{\lambda}{\mu}$$

$$\therefore L_q = \left[\frac{1}{(s-1)!}\left(\frac{\lambda}{\mu}\right)^s \frac{\lambda\mu}{(s\mu-\lambda)^2}\right] P_o$$

(ii) Expected number of customers in the system,

$$L_s = L_q + \frac{\lambda}{\mu}$$

(iii) Expected waiting time of a customer in the queue

$$W_q = \frac{L_q}{\lambda} = \left\{\frac{1}{(s-1)!}\left(\frac{\lambda}{\mu}\right)^s \left[\frac{\mu}{(s\mu-\lambda)^2}\right]\right\} P_o$$

(iv) Expected waiting time that a customer spends in the system

$$W_s = W_q + \frac{1}{\mu} = \frac{L_q}{\lambda} + \frac{1}{\mu}$$

(v) Probability that an arriving customer has to wait (busy period)

$$P(n \geq s) = \sum_{n=s}^{\infty} P_n = \sum_{n=s}^{\infty} \frac{1}{s!\,s^{n-s}}\left(\frac{\lambda}{\mu}\right)^n P_o$$

$$= \frac{1}{s!}\left(\frac{\lambda}{\mu}\right)^s P_o \sum_{n=0}^{\infty}\left(\frac{\lambda}{\mu}\right)^n = \frac{1}{s!}\left(\frac{\lambda}{\mu}\right)^s P_o \frac{1}{1-\left(\frac{\lambda}{\mu}\right)}$$

(vi) Since to sum to infinity in a geometric progression $s_\infty = \frac{r}{1-r}$

The probability that system shall be idle is

$$P_o = \left[\sum_{n=0}^{s-1} \frac{(s\rho)^n}{n!} + \frac{1}{s!}\frac{(s\rho)^s}{1-\rho}\right]^{-1}, \ \rho = \frac{\lambda}{s\mu}$$

$$= \left[\sum_{n=0}^{s-1} \frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^n + \frac{1}{s!}\left(\frac{\lambda}{\mu}\right)^n \frac{s\mu}{s\mu-\lambda}\right]^{-1}$$

Worked Out Examples

EXAMPLE 13.15

A telephone exchange has two long distance operators. The telephone company finds that during the break load, long distance calls arrive in a Poisson fashion at an average rate of 15 per hour. The length of service on these calls is approximately exponentially distributed with mean length of 5 minutes.

(i) What is the probability that a subscriber will have to wait for his long distance call during the peak hours of the day?

(ii) If subscribers will wait and are serviced in turn, what is the expected waiting time?

Solution: Average rate of arrivals = 15 per hour

$$\lambda = \frac{15}{60} = \frac{1}{4}$$

Average rate of departures = 5 minutes

$$\mu = \frac{1}{5}$$

Exchange has two servers, $s = 2$

$$\rho = \frac{\lambda}{s\mu} = \frac{5}{4.2} = \frac{5}{8}$$

(i) Probability that subscriber has to wait during the peak hours

$$P(n=2) = \frac{1}{s!}\left(\frac{\lambda}{s}\right)^s \frac{s\mu}{s\mu - \lambda} P_o$$

$$P_o = \left[\sum_{n=0}^{1} \frac{\left(2\left(\frac{5}{8}\right)\right)^n}{n!} + \frac{1}{2!}\frac{\left(\frac{5}{4}\right)^2}{\frac{3}{8}}\right]^{-1}$$

$$= \left[1 + \frac{5}{4} + \frac{1}{2}\cdot\frac{25}{16}\times\frac{8}{3}\right]^{-1}$$

$$= [4.33]^{-1} = 0.2307$$

$$P(n=2) = \left[\frac{1}{2!}\left(\frac{5}{4}\right)^2\left(\frac{\frac{2}{5}}{\frac{2}{5}-\frac{1}{4}}\right)\right]0.2307$$

$$= [0.78125(2.666)]0.2307$$

$$= 0.4806$$

(ii) Expected waiting time of a customer in the queue

$$W_q = \frac{1}{(s-1)!}\left(\frac{\lambda}{\mu}\right)^s \cdot \frac{\mu}{(s\mu - \lambda)^2} \cdot P_o$$

$$W_q = \left[\frac{1}{1!}\left(\frac{5}{4}\right)^2 \cdot \frac{\frac{1}{5}}{\left(\frac{2}{5}-\frac{1}{4}\right)^2}\right] 0.2307$$

$$= \frac{0.3125}{0.0225} \times 0.2307 = 3.204 \text{ minutes}$$

EXAMPLE 13.16

An insurance company has three claim adjusters in its branch office. People with claims against the company are found to arrive in a Poisson fashion, at an average rate of 20 per $\frac{8 \text{ hour}}{\text{day}}$. The amount of time that an adjuster spends with a claimant is found to have exponential distribution with mean service time of 40 minutes. Claimants are processed in the order of their appearance.

(i) How many hours a week can an adjuster expect to spent with claimants?

(ii) How much time, on the average does a claimant spend in the branch office?

Solution: Average rate of arrivals, $\lambda = \frac{20}{8} = \frac{5}{2}$

Average rate of service, $\mu = \frac{60}{40} = \frac{3}{2}$

Here, no. of services, $s = 3$

$$\rho = \frac{\lambda}{s\mu} = \frac{5}{2 \times \frac{3}{2} \times 3} = \frac{5}{9}$$

$$P_o = \left[\sum_{n=0}^{s-1} \frac{(s\rho)^n}{n!} + \frac{1}{s!}\frac{(s\rho)^s}{1-\rho}\right]^{-1}$$

$$= \left[\sum_{n=0}^{3} \frac{\left(3 \times \frac{5}{9}\right)^n}{n!} + \frac{1}{3!}\frac{\left(3 \times \frac{5}{9}\right)^3}{1-\frac{5}{9}}\right]^{-1}$$

$$= \left[1 + \frac{5}{3} + \left(\frac{5}{3}\right)^2 \cdot \frac{1}{2!} + \left(\frac{5}{3}\right)^3 \cdot \frac{1}{3!} + \frac{1}{3!}\frac{\left(\frac{5}{3}\right)^3}{\frac{4}{9}}\right]^{-1}$$

$$= \left[1 + \frac{5}{3} + 1.3888 + 0.7716 + 1.736\right]^{-1}$$

$$= \frac{1}{6.563}$$

$$= 0.152$$

$$P_n = \frac{1}{s!\,s^{n-s}}\left(\frac{\lambda}{\mu}\right)^n P_o$$

$$P_1 = \frac{1}{3!3^{1-3}}\left(\frac{\frac{5}{2}}{\frac{3}{2}}\right)^1\left(\frac{24}{139}\right)$$

$$P_2 = \frac{1}{3!3^{2-3}}\left(\frac{\frac{5}{2}}{\frac{3}{2}}\right)^2 (0.17266)$$

Hence expected number of adjusters at any specified time will be $3P_o + 2P_1 + P_2 = \frac{4}{3}$.

Probability that adjuster is idle $= \frac{4}{(3\times 3)} = \frac{4}{9}$

Probability that adjuster is busy $= 1 - \frac{4}{9} = \frac{5}{9}$

(i) Number of hours that an adjuster is expected to spend with claimants per week $= \frac{5}{9}\times 8(5)$

(ii) On the average, the time live spent by a claimant in the branch office

$$= W_s = \frac{L_q}{\lambda} + \frac{1}{\mu}$$

$$L_q = \sum_{n=s}^{\infty}(n-s)\frac{\rho^n}{s^{n-s}s!}P_o$$

$$= \sum_{n=3}^{\infty}(n-3)\frac{\left(\frac{5}{9}\right)^n}{3^{n-3}3!}(0.152)$$

$$= \frac{0.152}{6}\left[0 + \frac{1\left(\frac{5}{9}\right)^4}{3^1} + \frac{2\left(\frac{5}{9}\right)^5}{3^2} + \frac{3\left(\frac{5}{9}\right)^6}{3^3}\right] + \ldots$$

$$= \frac{(0.152)}{6}\frac{\left(\frac{5}{9}\right)^4}{3}\left[1 + 2\left(\frac{5}{9}\right)\cdot\frac{1}{3} + 3\left(\frac{5}{9}\cdot\frac{1}{3}\right)^2 + \ldots\right]$$

$$= \frac{0.152}{6}\frac{\left(\frac{5}{9}\right)^4}{3}\left(1 - \frac{5}{9}\cdot\frac{1}{3}\right)^{-2}$$

$$= 0.001211$$

$$W_s = \frac{0.001211}{\frac{5}{2}} + \frac{1}{\frac{3}{2}} = 0.667$$

Model V: {(m/m/s): (N/FCFS)}

- This is an extension of model IV with only difference in the waiting area for customers which are limited.
- Once the waiting area is full, the arriving customers are turned away and may or may not return.
- While performing economic analysis, cost associated with losing a customer is also taken into consideration along with cost per server and the cost of waiting.

Results:

(i) Probability P_o that a system shall be idle is

$$P_o = \sum_{n=0}^{s-1} \frac{(s\rho)^n}{n!} + \frac{1}{s!}\left(\frac{\lambda}{\mu}\right)^s \frac{s\mu}{s\mu - \lambda}\left\{1 - \left(\frac{\lambda}{s\mu}\right)^{N-s+1}\right\}\Bigg]^{-1}$$

where $\rho = \dfrac{\lambda}{s\mu} = 1$

$$\therefore P_o = \left[\sum_{n=0}^{s-1} \frac{(s\rho)^n}{n!} + \frac{1}{s!}\left(\frac{\lambda}{\mu}\right)^s (N - s + 1)\right]^{-1}$$

Here, for $N \to \infty$ and $\dfrac{\lambda}{s\mu} < 1,$ this result corresponds to that in model IV and for $s = 1$ this reduces to that in model III.

(ii) Effective arrival rate $\lambda_e = \lambda (1 - P_N)$

Effective traffic intensity, $\rho_e = \dfrac{\lambda_e}{\mu}$

(iii) Expected number of customers in the queue

$$L_q = \sum_{n=s}^{N} (n-s) P_n = \sum_{n=s}^{N} (n-s) \frac{s^n}{s!\, s^{n-s}} \left(\frac{\lambda}{s\mu}\right)^n P_o$$

$$= \sum_{n=s}^{N} (n-s) \frac{(s\rho)^n}{s!\, s^{n-s}} P_o = \frac{(s\rho)^s P_o \rho^{n-s}}{s!} \sum_{x=0}^{n-s} x\rho^{x-1}$$

where $x = n - s,\ \rho = \dfrac{\lambda}{s\mu}$

$$\therefore L_q = \frac{(s\rho)^3 \rho P_o}{s!} \sum_{x=0}^{N-s} \frac{d}{d\rho}(\rho^x)$$

$$= \frac{(s\rho)^s \rho}{s!} \frac{d}{d\rho}\left[\sum_{x=0}^{N-s} \rho^x\right] P_o$$

$$= \frac{(s\rho)^s \rho}{s!} \frac{d}{d\rho}[1 + \rho + \rho^2 + \cdots \rho^{N-s}]P_o$$

$$= \frac{(s\rho)^s \rho}{s!} \frac{d}{d\rho}\left[\frac{1-\rho^{N-S+1}}{1-\rho}\right]P_o$$

Since this is a GP with $N - S + 1$ terms,

$$\therefore L_q = \frac{(s\rho)^s \rho}{(1-\rho)^2 s!}[1 - \rho^{N-S+1} - (1-\rho)(N-S+1)\rho^{N-s}]P_o$$

$$\therefore L_q = \frac{(s\rho)^s \rho}{s!(1-\rho)^2}[1 - \rho^{N-S+1} - (1-\rho)(N-s+1)\rho^{N-s}]P_o$$

(iv) Expected number of customers in the system,

$$L_s = L_q + \left(\frac{\lambda}{\mu}\right)(1 - P_N)$$

$$= L_q + S - P_o \sum_{n=0}^{s-1} \frac{(s-n)}{n!}\left(\frac{\lambda}{\mu}\right)^n$$

(v) Expected waiting time in the system is

$$W_s = \frac{L_s}{\lambda(1-P_N)}$$

(vi) Expected waiting time in the queue is

$$W_q = W_s - \frac{1}{\mu} = \frac{L_q}{\lambda(1-P_N)}$$

(vii) Fraction server idle time

$$= 1 - \frac{L_s - L_q}{s} = 1 - \frac{\rho_e}{s}$$

If no queue is allowed then number of customers intend to join a queuing system which should not exceed the number of serves (i. e.,) $n \le s$.

$$\therefore P_n = \frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^n P_o \text{ and } P_o = \left[\sum_{n=0}^{s} \frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^n\right]^{-1}$$

Since no queue forms, $L_q = W_q = 0$.

The following are the results for the above case

(i) Fraction of potential customer loss, $P_s = \frac{1}{s!}\left(\frac{\lambda}{\mu}\right)^s P_o$

(ii) Effective arrival rate, $\lambda_{eff} = \lambda(1 - P_s)$

(iii) Effective waiting time in the system, $W_s = \dfrac{1}{\mu}$

(iv) Expected number of customers in the system, $L_s = \lambda W_s = \dfrac{\lambda}{\mu}$

(v) Fraction idle time for server $= 1 - \dfrac{\rho_{eff}}{s}$

EXAMPLE 13.17

A car servicing station has two bays, where service can be offered simultaneously. Due to space limitation, only four cars are accepted for servicing. The arrival pattern is Poisson with 120 cars per day. The service time in both ways is exponentially distributed with $\mu = 96$ cars per day per bay.

(i) Find the average number of cars waiting to be serviced.

(ii) Find the average number of cars in the service station.

(iii) Find the average time a car spends in the system.

Solution: Average rate of arrivals = 120 cars per day

$$\lambda = \frac{120}{24} = 5 \text{ per hour}$$

Average rate of services, $\mu = 96$ cars per day $\mu = \dfrac{96}{24} = 4$ per hour

Number of serves, $s = 2$
Maximum Number of cars, $N = 4$

(i) Average number of cars waiting to be serviced $= L_q$

$$P_o = \left[\sum_{n=0}^{s-1} \frac{(\lambda)^n}{n!(\mu)^n} + \frac{1}{s!}\left(\frac{\lambda}{\mu}\right)^s \frac{s\mu}{s\mu - \lambda}\left\{1 - \left(\frac{\lambda}{s\mu}\right)^{N-s+1}\right\}\right]^{-1}$$

$$= \left[1 + \frac{5}{4} + \frac{1}{2!}\left(\frac{5}{4}\right)^2 \frac{2(4)}{8-5}\left\{1 - \left(\frac{5}{8}\right)^{4-2+1}\right\}\right]^{-1}$$

$$= [1 + 1.25 + 2.0833(0.75586)]^{-1} = (3.8247)^{-1}$$

$$= 0.2614$$

(ii) Expected number of cars in the service station

$$L_s = L_q + s - P_o \sum_{n=0}^{s-1} \frac{(s-n)}{n!}\left(\frac{\lambda}{\mu}\right)^n$$

$$\rho = \frac{\lambda}{s\mu} = \frac{5}{2(4)} = \frac{5}{8}$$

$$L_q = \frac{(s\rho)^s \rho}{s!(1-\rho)^2}[1 - \rho^{N-s+1} - (1-\rho)(N-s+1)\rho^{N-s}]$$

$$= \frac{\left(2\times\frac{5}{8}\right)^2 \cdot \frac{5}{8}}{2!\left(1-\frac{5}{8}\right)^2}\left[1-\left(\frac{5}{8}\right)^{4-2+1} - \left(1-\frac{5}{8}\right)(4-2+1)\left(\frac{5}{8}\right)^{4-2}\right]P_o, \text{ where}$$

$$\therefore L_q = \left(\frac{5}{4}\right)^2 \frac{5}{8}\cdot\frac{1}{2}\left(\frac{8}{3}\right)^2 [1-0.244-0.4394]P_o$$

$$= 0.48828(7.11)\ 0.3165$$

$$L_q = (1.098)(0.2614)$$

$$= 0.2872$$

$$L_s = L_q + s - P_o \sum_{n=0}^{s-1} \frac{(s-n)}{n!}\left(\frac{\lambda}{\mu}\right)^n$$

$$= 0.2872 + 2 - 0.2614\left[\sum_{n=0}^{1} \frac{(2-n)}{n!}\left(\frac{\lambda}{\mu}\right)^n\right]$$

$$= 0.2872 + 2 - 0.2614\left[2\left(\frac{5}{4}\right)^0 + \frac{1}{1!}\left(\frac{5}{4}\right)\right]$$

$$= 2.2872 - 0.32675 = 1.96045 \simeq 2$$

$\therefore$ Expected number of cars in the service station = 2

(iii) Average time a car spends in the system

$$W_q = W_s - \frac{1}{\mu} = \frac{L_q}{\lambda(1-P_N)}$$

$$P_N(n \geq s) = \frac{1}{s!}\left(\frac{\lambda}{\mu}\right)^s \left(\frac{s\mu}{s\mu-\lambda}\right)P_o$$

$$= \frac{1}{2!}\left(\frac{5}{4}\right)^2 \left(\frac{2\times 4}{2\times 4-5}\right)0.2614 = \left(\frac{25}{2\times 16}\right)\left(\frac{8}{3}\right)(0.2614)$$

$$P_N = 0.5445$$

$$\therefore W_q = \frac{0.2872}{5(1-0.5445)} = \frac{0.2872}{2.2775} = 0.126 \text{ hr}$$

Work Book Exercises

13. A steel fabrication plant was considering the installation of a second tool crib in the plant to save walking time of the skilled craftsman who checkout equipment at the tool cribs. The Poisson/exponential assumptions about arrivals are justified in this case. The time of the craftsman is valued at $\frac{₹20}{\text{hour}}$. The current facility receives an average of 10 calls per hour, with two cribs, each would have an average of five calls per hour. Currently there are two attendants, each of them services one craftsman per hour. Each could do just as well in a separate tool crib. There would be added

average inventory costs over the year of $\frac{₹20}{\text{hour}}$ with the separate tool cribs. However, each craftsman would require 6 minutes less walking time per call. Evaluate the proposal to set up a new crib so that each attendant would run one crib.

14. A telephone exchange has two long distance operators. The telephone company finds that during the peak load, long distance calls arrive in a Poisson fashion at an average rate of 15 per hour. The length of service on these calls is approximately exponentially distributed with mean length of 5 minutes.

    (i) What is the probability that a subscriber will have to wait for his long distance call during the peak hours of the day?

    (ii) If subscribers will wait and are serviced in turn, what is the expected waiting time?

[**Ans.:** $\lambda = \frac{1}{4},\ \mu = \frac{1}{5},\ s = 2,\ \rho = \frac{\lambda}{s\mu} = \frac{5}{8}$

$$P_o = \left[\sum_{n=0}^{s-1} \frac{(s\rho)^n}{n!} + \frac{1}{s!}\left(\frac{(s\rho)^s}{1-\rho}\right)\right]^{-1} = \frac{3}{13}$$

$$P\left(n = 2 = \frac{1}{s!}\left(\frac{\lambda}{\mu}\right)^s \frac{s\mu}{s\mu - \lambda}\right) P_o = 0.48$$

$$W_q = \frac{1}{(s-1)!}\left(\frac{\lambda}{\mu}\right)^s \frac{\mu}{(s\mu - \lambda)^2} P_o = 3.2 \text{ minutes }]$$

15. A bank has two tellers working on savings accounts. The first teller handles withdrawals only while the second teller handles deposits only. It has been found that the service time distribution for the deposits and withdrawals both is exponential with mean service time 3 minutes per customer. Depositors are found to arrive in Poisson fashion throughout the day with mean arrival rate 16 per hour. Withdrawers also arrive in a Poisson fashion with mean arrival rate 14 per hour.

    (i) What would be the effect on the average waiting time for depositors and withdrawers if each teller could handle both withdrawals and deposits?

    (ii) What would be the effect if this could only be accomplished by increasing the service time to 3.5 minutes?

[**Ans. :** $\mu = \frac{1}{3}$ per minute, $\lambda_1 = \frac{16}{\text{hour}}$, $\lambda_2 = \frac{14}{\text{hour}}$

$$= \frac{20}{\text{hour}}$$

$$W_{q_1} = \frac{1}{\mu}\frac{\lambda_1}{\mu - \lambda_1} = 12 \text{ minutes},\ W_{q_2} = 7 \text{ minutes}]$$

(i) $\lambda = \lambda_1 + \lambda_2 = \frac{30}{\text{hour}},\ \mu = \frac{20}{\text{hour}},\ \frac{\lambda}{\mu} = \frac{3}{2},\ s = 2$

$$P_o = \left[\sum_{n=0}^{s-1} \frac{\left(\frac{\lambda}{\mu}\right)^n}{n!} + \frac{\left(\frac{\lambda}{\mu}\right)^s}{s!} \cdot \frac{s\mu}{s\mu - \lambda}\right]^{-1} = \frac{1}{7}$$

$$W_q = \frac{\mu\left(\frac{\lambda}{\mu}\right)^s}{(s-1)^1(s\mu-\lambda)^2} \quad P_o = \frac{27}{7} = 3.86 \text{ minutes}$$

(ii) $\mu = \frac{\frac{1}{3.5}}{\text{minutes}} = \frac{\frac{120}{7}}{\text{hour}}, \lambda = \frac{30}{\text{hour}}, \frac{\lambda}{\mu} = \frac{7}{4}$

$$s = 2$$

$$P_o = \frac{1}{15}, W_q = \frac{343}{1800} \text{ hours} = 11.43 \text{ minutes}$$

16. A company has two tool cribs each having a single clerk in its manufacturing area. One tool crib handles only the tools for the heavy machinery, while the second one handles all other tools. It is observed that for each tool crib the arrivals follow a Poisson distribution with a mean of 20 per hour, and the service time distribution is negative exponential with a mean of 2 minutes.

The tool manager feels that if tool cribs are combined in such a way that either clerk can handle any kind of tool as demand arises would be more efficient and the waiting problem could be reduced to some extent. It is believed that the mean arrival rate at the two tool cribs will be 40 per hour; while the service time remains unchanged.

(i) Compare the status of the queue and the proposal with respect to the total expected number of machines at the tool cribs.

(ii) Calculate the expected waiting time including service time for each mechanic

(iii) Find the probability that he has to wait for more than 5 minutes.

[**Ans.:** (i) When tool crib works independently $\lambda = \frac{20}{\text{hour}}, \mu = \frac{2}{\text{minute}} = \frac{30}{\text{hour}}$

$$\rho = \frac{\lambda}{\mu} = \frac{2}{3}, L_s = \frac{\rho}{1-\rho} = 2, W_s = \frac{L_s}{\lambda} = 6 \text{ minutes (each)}$$

(ii) $W_s = \frac{L_s}{\lambda} = 3.6$ minutes.

(iii) $P(t > 5) = e^{-\mu(1-\rho)t}$

$$= 0.435$$

When both tool cribs work jointly: $\lambda = \frac{40}{\text{hour}}, \mu = \frac{30}{\text{hour}}$

$$s = 2 \quad \rho = \frac{\lambda}{s\mu} = \frac{2}{3}, \quad P_o = \left[\sum_{n=0}^{s-1} \frac{(s\rho)^n}{n!} + \frac{1}{s!}\frac{(s\rho)^s}{1-\rho}\right]^{-1} = \frac{1}{5}$$

$$L_s = L_q + \frac{\lambda}{\mu} = 2.4 \text{ persons}]$$

DEFINITIONS AT A GLANCE

Arrival Pattern: The way in which a queue is formed is known as an arrival pattern.

Inter-arrival Time: The time between two successive customers is called inter-arrival time.

Service Pattern: The way in which the server serves.

Queue Size: Number of customers in a queue.

Queue Discipline: The order in which the customers are selected in a queue.

Queue Behaviour: The behaviour of customers in a queue is its queue behaviour.

Jockeying: A customer in the middle of a queue joins another queue with the idea of reducing the waiting time.

Balking: A customer sometimes may not enter a queue altogether because he may anticipate long waiting.

Reneging: A customer who has been waiting for long time may leave the queue.

Queue Length: The number of customers being served in addition to the line length.

Pure Birth Process: The arrival process in which it is assumed that customers arrive at the queuing system and never leaves it.

FORMULAE AT A GLANCE

Model I {m/m/1}: {∞/ FCFS}:

- Expected no. of customers in the system

$$L_s = \frac{\rho}{(1-\rho)} \quad \text{where } \rho = \frac{\lambda}{\mu}$$

$$= \frac{\lambda}{\mu - \lambda}$$

- Expected number of customers waiting in the queue

$$L_q = \frac{\lambda^2}{\mu(\mu - \lambda)}$$

- Expected waiting time for a customer in the queue

$$W_q = \frac{L_q}{\lambda}$$

- Expected waiting time for a customer in the system (waiting and service) is

$$W_s = \frac{L_s}{\lambda}$$

- Probability that number of customers in the system is greater than or equal to k

$$P(n \geq k) = \left(\frac{\lambda}{\mu}\right)^k$$

- Variance of queue length

$$\text{var}(n) = \frac{\lambda\mu}{(\mu - \lambda^2}$$

- Probability that queue is non-empty is

$$P(n > 1) = \left(\frac{\lambda}{\mu}\right)^2$$

- Expected length of non-empty queue is

$$L_b = \frac{\mu}{\mu - \lambda}$$

- Probability of an arrival during the service time when system contains *r* customers.

$$P(n = r) = \left(\frac{\lambda}{\lambda + \mu}\right)^r \left(\frac{\mu}{\lambda + \mu}\right)$$

Model II: {(m/m/1): (∞/SIRO)}

$$P_n = (1-\rho)\,\rho^n$$

Model III: {(m/m/1): (N/FCFS)}

- Expected number of customers in the system

$$L_s = \begin{cases} \dfrac{\rho}{1-\rho} - \dfrac{(N+1)\rho^{N+1}}{1-\rho^{N+1}}, & \rho \neq 1 \\ \dfrac{N}{2}, & \rho = 1, \quad \lambda = \mu \end{cases}$$

- Expected number of customers waiting in the system,

$$L_q = L_s - \frac{\lambda(1-P_N)}{\mu}$$

- Expected waiting time of a customer in the system (waiting + service),

$$W_s = \frac{L_s}{\lambda(1-P_N)}$$

- Expected waiting time of a customer in the queue

$$W_q = \frac{L_q}{\lambda(1-P_N)}$$

- Fraction of time the system is full

$$P_N = P_o \rho^N$$

- Effective arrival rate, $\lambda\rho = \lambda\,(1 - P_N)$
- Effective traffic intensity, $\rho_e = \dfrac{\lambda_e}{\mu}$

Model IV: {(m/m/s): (∞, FCFS)}

- Expected number of customers waiting in the queue

$$L_q = \sum_{n=s}^{\infty} (n-s) \frac{\rho^n}{s^{n-s} s!} P_o$$

- Expected number of customers in the system

$$L_s = L_q + \frac{\lambda}{\mu}$$

- Expected waiting time of a customer in the system in the queue, $W_q = \frac{L_q}{\lambda}$
- Expected waiting time that a customer spends in the system, $W_s = \frac{L_q}{\lambda} + \frac{1}{\mu}$
- Probability that an arrival customer has to wait $P(n \geq s) = \frac{1}{s!}\left(\frac{\lambda}{\mu}\right)^s P_o \frac{1}{\left(1-\frac{\lambda}{\mu}\right)}$
- Probability that a system shall be idle

$$P_o = \left[\sum_{n=0}^{s-1} \frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^n + \frac{1}{s!}\left(\frac{\lambda}{\mu}\right)^n \left(\frac{s\mu}{s\mu-\lambda}\right)\right]^{-1}$$

Model V: {(m/m/s): (N/FCFS)}

- Probability that a system shall be idle

$$P_o = \left[\sum_{n=0}^{s-1} \frac{(s\rho)^n}{n!} + \frac{1}{s!}\left(\frac{\lambda}{\mu}\right)^s (N-S+1)\right]^{-1}$$

- Effective arrival rate

$$\lambda_e = \lambda(1-P_N)$$

- Effective traffic intensity,

$$\rho_e = \frac{\lambda_e}{\mu}$$

- Expected number of customers in the queue

$$L_q = \frac{(s\rho)^s \rho}{s!(1-e)^2}[1-\rho^{N-s+1} - (1-\rho)(N-S+1)\rho^{N-s}]P_o$$

- Expected number of customers in the system

$$L_s = L_q + S - P_o \sum_{n=0}^{s-1} \frac{(s-n)}{n!}\left(\frac{\lambda}{\mu}\right)^n$$

- Expected waiting time in the system is

$$W_s = \frac{L_s}{\lambda(1-P_N)}$$

- Expected waiting time in the queue

$$W_q = W_s - \frac{1}{\mu} = \frac{L_q}{\lambda(1-P_N)}$$

- Fraction server idle time

$$= 1 - \frac{\rho_e}{s}$$

- When $n \le s$, $P_n = \frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^n P_o$, where $P_o = \left[\sum_{n=0}^{s} \frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^n\right]^{-1}$

OBJECTIVE TYPE QUESTIONS

1. A customer in the middle of a queue when joins another queue with the idea of reducing the waiting time, then the queue behaviour is called _____.
   (a) Queue size (b) Reneging
   (c) Jockeying (d) none

2. A customer sometimes may not enter a queue altogether because he may anticipate long waiting, then the queue behaviour is called _____.
   (a) Jockeying (b) Balking
   (c) Reneging (d) none

3. A customer who has been waiting for a long time may leave the queue, then the queue behaviour is called _____.
   (a) Reneging (b) Balking
   (c) Jockeying (d) none

4. The arrival process in which it is assumed that the customers arrive at the queuing system and never leaves it, the process is called _____.
   (a) Pure birth process (b) Pure death process
   (c) both (a) and (b) (d) none

5. In the model I {(m/m/1)}: {∞/FCFS}, expected waiting time for a customer in the queue is _____.
   (a) $L_q = \frac{\lambda}{\mu(\mu-\lambda)}$ (b) $L_q = \frac{\lambda}{\mu-\lambda}$
   (c) $L_q = \frac{\lambda^2}{\mu(\mu-\lambda)}$ (d) none

6. In the model I {(m/m/1)}: {∞/FCFS}, probability that queue is non-empty is _____.
   (a) $P(n>1) = \left(\frac{\lambda}{\mu}\right)$ (b) $P(n>1) = \left(\frac{\lambda}{\mu}\right)^2$
   (c) $P(n>1) = \left(\frac{\lambda}{\mu}\right)^3$ (d) none

7. In the model II {(m/m/1): (∞/SIRO)}, expected number of customers waiting in the system is _____.

(a) $L_q = \frac{\lambda(1-P_N)}{\mu}$ (b) $L_q = \frac{L_S - P_N}{\mu}$

(c) $L_q = L_S - \frac{\lambda(1-P_N)}{\mu}$ (d) none

8. In the model II {(m/m/1): (∞/SIRO)}, effective arrival rate is _____.

(a) $\lambda\rho = \lambda(1-P_N)$ (b) $\lambda = 1-P_N$

(c) $\lambda = \frac{\lambda}{\rho} - P_N$ (d) none

9. In the model IV {(m/m/s): (∞/FCFS)}, expected waiting time of a customer in the system in the queue is _____.

(a) $W_q = \frac{L_s}{\lambda}$ (b) $W_q = (L_q)\lambda$

(c) $W_q = \frac{L_q}{\lambda}$ (d) none

10. In the model IV {(m/m/s): (∞/FCFS)}, probability that an arrival customer was to wait is _____.

(a) $P(n \geq s) = \frac{P_o}{1-\frac{\lambda}{\mu}}$ (b) $P(n \geq s) = \frac{1}{s!}\left(\frac{\lambda}{\mu}\right)^s \frac{P_o}{\left(1-\frac{\lambda}{\mu}\right)}$

(c) $P(n \geq s) = \left(\frac{\lambda}{\mu}\right)^s \frac{P_o}{\left(1-\frac{\lambda}{\mu}\right)}$ (d) none

11. In the model V {(m/m/s): (N/FCFS)}, expected waiting time in the queue is _____.

(a) $W_q = \frac{L_s}{1-P_N}$ (b) $W_q = \frac{L_q}{1-P_N}$

(c) $W_q = \frac{L_q}{\lambda(1-P_N)}$ (d) none

12. In the model V {(m/m/s): (N/FCFS)}, fraction server idle time is _____.

(a) $1-\frac{\rho_e}{s}$ (b) $\frac{\rho_e}{s}$

(c) $1-\frac{L_q}{s}$ (d) none

ANSWERS

1. (c) 2. (b) 3. (a) 4. (a) 5. (c) 6. (b) 7. (c) 8. (a)
9. (c) 10. (b) 11. (c) 12. (a)

14 Design of Experiments

Prerequisites

Before you start reading this unit, you should:

- Know the procedure used in testing hypothesis
- Computation of standard deviation and variance

Learning Objectives

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

- Compare more than two population means using analysis of variance
- Understand the assumption involved in analysis of variance technique
- Differentiate between one-way and two-way classifications
- Understand the analysis from decomposition of the individual observations
- Understand completely randomized and daily square design and randomized block design

INTRODUCTION

In many practical situations we are interested to determine—what are those variables which affect the particular dependent variable of the many independent variables?

The analysis of variance was developed by R. A. Fisher. This is at first used in agricultural research. However, now-a-days it is widely applied in manufacturing industries, quality control engineering to compare the production from different assembly lines or different plants.

14.1 ASSUMPTIONS OF ANALYSIS OF VARIANCE

The following are the assumptions made before actually carrying out the procedure:

(i) The data are quantitative in nature and are normally distributed. If the data is nominal or ordinal where results are given in ranks or percentages, Analysis of Variance (ANOVA) should not be carried out.

(ii) If the data are not exactly normally distributed but are close to a normal distribution, AVOVA can be used. The data are not highly skewed.

(iii) The samples are drawn randomly and independently, from the population. In case of an experimental design, the treatments must be assigned to test units by means of some randomizing device. In case this assumption does not hold, inferences drawn from ANOVA will not be valid.

(iv) The variances of the population from which the samples have been drawn are equal. If the sample sizes are all equal, then non-compliance of this assumption does not seriously affect the inferences based on F-distribution.

Suppose the effectiveness of these methods of teaching the programs of certain computer language is to be composed. Method *A* is a straight computer-based instruction, Method *B* involves the personal attention of an instructor and some direct experience working with the computer, and Method *C* involves the personal attention of an instructor but no work with the computer. An achievement test is conducted by taking random samples of size 5 from large groups of persons taught by the various methods.

Scores of achievement test

	Method *A*	Method *B*	Method *C*
	70	88	69
	74	78	74
	64	84	73
	68	82	76
	69	83	73
Total	345	415	365
Mean	69	83	73

The model for the above situation may be considered as follows:

There are five observations taken from each of three populations with means μ_1, μ_2, μ_3, respectively. Suppose we are interested to list the hypothesis that the means of these three populations are equal. Here we can attribute within the sample variation to chance or random variation. Our aim is to determine if the differences between the three sample means are due to random variation alone. In this section, we consider the statistical analysis of the one-way classification or completely randomized design.

Consider another example. Suppose that a manufacturer of plumbing materials wishes to compare the performance of several kinds of materials to be used in the underground water pipes. If physical conditions such as soil acidity, depth of pipe, and mineral content of water were all kept fixed, then which material is the best? The conclusions would be valid for the given set of conditions. However, the manufacturer wants to know which material is best over a wide variety of conditions. To design an experiment, it is advisable to specify that pipe of each material be buried at each of several depths in each of several kinds of soil, and in locations where the water varies in hardness.

14.2 ONE-WAY CLASSIFICATION

Suppose the experimenter has the results of *K* independent random samples from *K* different populations and we are interested in testing the hypothesis that the means of these *K* populations are equal.

Caution:

- *K* populations refer to the data concerning *K* treatments, *K* groups, etc.

Let y_{ij} denote the j^{th} observation in the i^{th} sample. Let $n_1, n_2, n_3, \ldots n_K$ denote the sizes of samples 1, 2, … *K* respectively.

Scheme for one-way classification

	Observations	Means	Sum of squares of deviations
Sample 1	$y_{11}, y_{12}, \ldots y_{1j}, \ldots y_{1n_1}$	$\bar{y}_1$	$\sum_{j=1}^{n_1}(y_{1j}-\bar{y}_1)^2$
Sample 2	$y_{21}, y_{22}, \ldots y_{2j}, \ldots y_{2n_2}$	$\bar{y}_2$	$\sum_{j=1}^{n_2}(y_{2j}-\bar{y}_2)^2$
⋮	⋮	⋮	⋮
Sample i	$y_{i1}, y_{i2}, \ldots y_{ij}, \ldots y_{in_i}$	$\bar{y}_i$	$\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_i)^2$
⋮	⋮	⋮	⋮
Sample K	$y_{K1}, y_{K2}, \ldots y_{Kj}, \ldots y_{Kn_K}$	$\bar{y}_K$	$\sum_{j=1}^{n_K}(y_{Kj}-\bar{y}_K)^2$

Where the means of the above sample are calculated as follows:

$$\bar{y}_1 = \frac{\sum_{j=1}^{n_1} y_{1j}}{n_1} \quad \cdots \quad \bar{y}_2 = \frac{\sum_{j=1}^{n_2} y_{2j}}{n_2}$$

$$\vdots \qquad\qquad \vdots$$

$$\bar{y}_i = \frac{\sum_{j=1}^{n_1} y_{ij}}{n_i} \quad \cdots \quad \bar{y}_K = \frac{\sum_{j=1}^{n_K} y_{Kj}}{n_K}$$

Let $\bar{T}$ denote the sum of all the observations and N denote total sample size.

$$\bar{T} = \sum_{i=1}^{K} \sum_{j=1}^{n_i} y_{ij} \text{ and } N = \sum_{i=1}^{K} n_i$$

Let $\bar{y}$ denote the overall sample mean.

$$\bar{y} = \frac{\bar{T}}{N} = \frac{\sum_{i=1}^{K} \sum_{j=1}^{n_i} y_{ij}}{\sum_{i=1}^{K} n_i}$$

$$= \frac{\sum_{i=1}^{K} \bar{y}_i n_i}{\sum_{i=1}^{K} n_i}$$

If we look at the previous example, with $K = 3$ treatments and equal sample sizes of each treatment, where y_{ij} is the j^{th} score obtained in an achievement test of i^{th} method. $\bar{y}_i$ denotes the mean of the measurements by the i^{th} method. $\bar{y}$ is the overall mean of all observations, $n = 15$.

Caution:

- $\bar{y}$ is also called grand mean of all observations.

	Method *A*	Method *B*	Method *C*
	70	88	69
	74	78	74
	64	84	73
	68	82	76
	69	83	73
Totals	345	415	365
Means $\bar{y}_i$	69	83	73

The total sample size $N = 5 + 5 + 5 = 15$ observations and the sum of all observations.

$$\bar{T} = 1125$$

The grand mean can be calculated as follows:

$$\bar{y} = \frac{\bar{T}}{N} = \frac{\sum_{i=1}^{K} \sum_{j=1}^{n:} y_{ij}}{N} = \frac{1125}{15} = 75$$

Now we shall prove a theorem which gives an identity for one-way ANOVA.

Theorem 14.1

Let there be K samples each of size $n_1, n_2, \ldots n_K$. Let y_{ij} denote j^{th} observation of the i^{th} sample. Then

$$\sum_{i=1}^{K} \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2 = \sum_{i=1}^{K} \sum_{i=1}^{n_i} (y_{ij} - \bar{y}_i)^2 + \sum_{i=1}^{K} n_i (\bar{y}_i - \bar{y})^2$$

Proof:

Consider $$y_{ij} - \bar{y} = (y_{ij} - \bar{y}_i) + (\bar{y}_i - \bar{y})$$

{Adding and subtractions the term $\bar{y}_i$}

Squaring on both the sides we get,

$$(y_{ij} - \bar{y})^2 = (y_{ij} - \bar{y}_i)^2 + (\bar{y}_i - \bar{y})^2 + 2(y_{ij} - \bar{y})(\bar{y}_i - \bar{y})$$

Summing the above equation over i and j on both the sides,

$$\sum_{i=1}^{K} \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2 = \sum_{i=1}^{K} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 + \sum_{i=1}^{K} \sum_{j=1}^{n_i} (\bar{y}_i - \bar{y})^2 + 2\sum_{i=1}^{K} \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2 (\bar{y}_i - \bar{y}) \quad (14.1)$$

Consider the third term of equation 14.1.

$$\sum_{i=1}^{K} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)(\bar{y}_i - \bar{y}) = \sum_{i=1}^{K} (\bar{y}_i - \bar{y}) \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)$$

$$\sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i) = (y_{i1} - \bar{y}_i) + (y_{i2} - \bar{y}_i) + \ldots (y_{in_i} - \bar{y}_i)$$

$$= (y_{i1} + y_{i2} + \cdots + y_{ini}) - n_i \bar{y}_i$$

$$= n_i \bar{y}_i - n_i \bar{y}_i = 0$$

$$= \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i) = 0$$

Consider the second term of equation 14.1.

$$\sum_{i=1}^{K} \sum_{j=1}^{n_i} (\bar{y}_i - \bar{y})^2 = \sum_{i=1}^{K} n_i (\bar{y}_i - \bar{y})^2$$

Since the summand does not involve subscript *j*.

∴ Equation (14.1) becomes

$$\sum_{i=1}^{K} \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2 = \sum_{i=1}^{K} \sum_{j=1}^{n_i} (\bar{y}_{ij} - \bar{y}_i)^2 + \sum_{i=1}^{K} n_i (\bar{y}_i - \bar{y})^2$$

$$SST = SSE + SS(Tr)$$

where SST is called Total Sum of Squares, *SSE* is called Error Sum of Squares which estimates random error and *SS*(*Tr*) is Treatment Sum of Squares.

Let μ_i denote the mean of i^{th} population and σ^2 denote the common variance of the *K* populations. Then each observation y_{ij} can be expressed as

$$y_{ij} = \mu_i + \in_{ij}, i = 1, 2, \ldots K, \; j = 1, 2, \ldots n_i \tag{14.2}$$

where the term $\in_{ij}$ represents that the random error are independent, normally distributed with zero means and common variance σ^2.

Let μ_i be replaced with $\mu + \alpha_i$ where μ is the mean of all μ_i in the experiment and α_i is the effect of i^{th} treatment.

$$\therefore \mu = \frac{\sum_{i=1}^{K} n_i \mu_i}{N}$$

Hence, $\sum_{i=1}^{K} n_i \alpha_i = 0$.

Substituting these in equation (14.2) we get,

$$y_{ij} = \mu + \alpha_i + \in_{ij} \qquad i = 1, 2, \ldots K$$

$$j = 1, 2, \ldots n_i$$

This called the model equation for one-way classification.

To test the null hypothesis that the *K* population means are all equal, we compare two estimates of σ^2:

(i) Based on variation among the sample means

(ii) Based on variation within the samples

A test for the equality of treatment means is performed. When the null hypothesis is false, the treatment sample mean square can be expected to exceed the error sample mean square. If the null hypothesis is true, the two mean squares are independent and we have *F*-ratio for treatments given by

$$F = \frac{\left(\frac{SS(Tr)}{K-1}\right)}{\left(\frac{SSE}{N-K}\right)} = \frac{\left[\frac{\sum_{i=1}^{K} n_i(\bar{y}_i - \bar{y})^2}{K-1}\right]}{\left[\frac{\sum_{i=1}^{K}\sum_{j=1}^{n_i}(\bar{y}_{ij} - \bar{y}_i)^2}{N-K}\right]}$$

Follows F-distribution with $(K-1, N-K)$ degrees of freedom.

If $F > F_{\alpha}$, null hypothesis will be rejected at α level of significance. A large value of F indicates large differences between the sample means.

Caution:

- When the population means are equal,

  treatment mean square $= \frac{\sum_{i=1}^{K} n_i(\bar{y}_i - \bar{y})^2}{K-1}$ and Error mean square $= \frac{\sum_{i=1}^{K}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2}{N-K}$ are estimates of σ^2.

- Each sum of squares is first converted to a mean square to test for equality of treatment means.

Hence, the 'ANOVA' can be summarized as follows, in the form of a table:

One-way ANOVA table

Source of variation	Sum of squares	Degrees of freedom	Mean square	F
Treatments	$SS(Tr)$	$K-1$	$MS(Tr) = \frac{SS(Tr)}{(K-1)}$	$F = \frac{MS(Tr)}{MSE}$
Error	SSE	$N-K$	$MSE = \frac{SSE}{N-K}$	
Total	$N-1$	SST		

While calculating F formula, the following shortcut method can be adopted:

$$SSE = SST - SS(Tr)$$

$$SST = \sum_{i=1}^{K}\sum_{j=1}^{n_i} y_{ij}^2 - C$$

$$SS(Tr) = \sum_{i=1}^{K} \frac{T_i^2}{n_i} - C$$

where C is called correction term for the mean is given by $C = \frac{\bar{T}^2}{N}$

$$N = \sum_{i=1}^{K} n_i, \bar{T} = \sum_{i=1}^{K} T_i \text{ and } T_i = \sum_{j=1}^{n_i} y_{ij}$$

T_i is total of n_i observations in the i^{th} sample.

Worked Out Examples

EXAMPLE 14.1

The following are the numbers of mistakes made in 4 successive days for 4 technicians working for a photographic laboratory.

	Technician I	Technician II	Technician III	Technician IV
	5	13	9	8
	13	8	11	11
	9	11	6	7
	10	13	10	10
Total	37	45	36	36

Perform an ANOVA to test at 0.05 level of significance whether the difference among the four sample means can be attributed to chance.

Solution:

(i) **Null hypothesis:** $\mu_1 = \mu_2 = \mu_3 = \mu_4$

(ii) **Alternative hypothesis:** H_1: μ_i's are not all equal

(iii) **Level of Significance:** $\alpha = 0.05$

(iv) **Criterion:** Reject if $F_{cal} > F_{\alpha}$ at α level of significance where $F = F_{cal}$ is to be determined from analysis of variables.

(v) **Calculation:**

The tables for the 4 samples each of size 4 are 37, 45, 36, and 36
Grand Total $\overline{T} = 154$
Total sample size = 4 + 4 + 4 + 4 = 16

Correction term for the mean $C = \dfrac{\overline{T}^2}{N} = \dfrac{(154)^2}{16}$

$= 1482.25$

$$SST = \sum_{i=1}^{K} \sum_{j=1}^{n_i} y_{ij}^2 - C$$

$= [5^2 + 13^2 + 9^2 + 10^2 + 13^2 + 8^2 + ... + 10^2] - 1482.25$

$= 1570 - 1482.25$

$= 87.75$

Treatment sum of squares $SS(Tr) = \sum_{i=1}^{K} \dfrac{T_i^2}{n_i} - C$

$= \left[\dfrac{37^2}{4} + \dfrac{45^2}{4} + \dfrac{36^2}{4} + \dfrac{36^2}{4}\right] - 1482.25$

$= [342.25 + 506.25 + 324 + 324] - 1482.25$

$$= 1496.5 - 1482.25$$
$$= 14.25$$

Error sum of squares, $SSE = SST - SS(Tr)$
$$= 87.75 - 14.25$$
$$= 73.5$$

Now table for ANOVA can be obtained as follows:

Source of variation	Sum of squares	Degree of freedom	Mean squares	F
Technicians	$SS(Tr) = 14.35$	$K - 1 = 3$	$\frac{14.25}{3} = 4.75$	$F = \frac{4.75}{6.125}$
Error	$SSE = 73.5$	$N - K = 12$	$\frac{73.5}{12} = 6.125$	0.7755

The calculated value of $F = 0.7755$. The tabulated value of F namely F_{α} for (3, 12) degrees of freedom at $\alpha = 0.05$ level of significance = 3.49.

(vi) **Decision:** Since the calculated value of F + tabulated value F_{α} we accept null hypothesis. Hence there are no significant differences among the four sample means.

EXAMPLE 14.2

Samples of peanut butter produced by two different manufacturers are tested for aflatoxin content with the following results:

Aflatoxin content (*ppb*)	
Brand *A*	Brand *B*
0.5	4.7
0.0	6.2
3.2	0.0
1.4	10.5
0.0	2.1
1.0	0.8
8.6	2.9

(a) Use ANOVA to test whether the two brands differ in aflatoxin content.
(b) Test the same hypothesis using a two-sample test.
(c) Show that t-statistic with γ degrees of freedom and the F-statistic with $(1, \gamma)$ degrees of freedom are related by the former $F(1, \gamma) = t^2(\gamma)$.

Solution:

(a) The totals of the two samples each of size 7 are 14.7 and 27.2, respectively.

Grand total $\bar{T} = 41.9$

Total sample size, $N = 7 + 7 = 14$

Correction term for the mean $C = \dfrac{\bar{T}^2}{N} = \dfrac{(41.9)^2}{14}$

$= 125.4007$

Total sum of squares, $SST = \sum_{i=1}^{K} \sum_{j=1}^{n_i} y_{ij}^{\ 2} - C$

$= [0.5^2 + 3.2^2 + 1.4^2 + 1^2 + 8.6^2 + 4.7^2 + ... \ 2.9^2] - 125.4007$

$= 271.65 - 125.4007$

$= 146.2493$

Treatment sum of squares, $SS(Tr) = \sum_{i=1}^{K} \dfrac{T_i^2}{n_i} - C$

$= \left[\dfrac{14.7^2}{7} + \dfrac{27.2^2}{7}\right] - 125.4007$

$= 136.5614 - 125.4007$

$= 11.1607$

Error sum of squares, $SSE = SST - SS(Tr)$

$= 146.2493 - 11.1607$

$= 135.0886$

The table of ANOVA is as follows:

Source of variation	Sum of squares	Degrees of freedom	Mean square	F
Brands	$SS(Tr) = 11.1607$	$K - 1 = 2 - 1 = 1$	$\dfrac{11.1607}{1} = 11.1607$	$F = \dfrac{11.1607}{11.2573}$
Error	$SSE = 135.0886$	$N - K = 14 - 2 = 12$	$\dfrac{135.0886}{12} = 11.257$	$= 0.9914$

(i) **Null hypothesis:** H_0: $\mu_1 = \mu_2$

(ii) **Alternative hypothesis:** H_1: $\mu_1 \neq \mu_2$

(iii) **Level of significance:** $\alpha = 0.05$

(iv) **Criterion:** Reject null hypothesis if $F > F_\alpha$ at α level of significance

Hence the calculated value of $F = 0.9914$ Tabulated value of $F = F_\alpha = 0.05$ level of significance $F_\alpha = 4.75$.

(v) **Decision:** Since the calculated value of F is not greater than tabulated value of F, we accept Null Hypothesis at 5% level of significance. Hence the two brands do not differ in aflatoxin content.

(b)

Brand *A*			Brand *B*		
x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	y_i	$y_i - \bar{y}$	$(y_i - \bar{y})^2$
0.5	–1.6	2.56	4.7	0.82	0.6724
0.0	–2.1	4.41	6.2	2.32	5.3824
3.2	1.1	1.21	0.0	–3.88	15.0544
1.4	–0.7	0.49	10.5	6.62	43.8244
0.0	–2.1	4.41	2.1	1.78	3.1684
1.0	–1.1	1.21	0.8	–3.08	9.4864
8.6	6.5	42.25	2.9	–0.98	0.9604
$\sum x_i = 14.7$	$\sum (x_i - \bar{x})^2 = 56.54$		$\sum y_i = 27.2$	$\sum (y_i - \bar{y})^2 = 78.5488$	

$$\sum x_i = 14.7$$

$$\bar{x} = \frac{\sum x_i}{n_1} = \frac{14.7}{7}$$

$$= 2.1$$

$$\sum y_i = 27.2$$

$$\bar{y} = \frac{\sum y_i}{n_2} = \frac{27.2}{7}$$

$$= 3.88$$

The *t*-test can be used to test for equality of means.

$$\text{Under } H_o,\ t = \frac{\bar{x} - \bar{y}}{\sqrt{S^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim t_{n_1 + n_2 - 2}$$

Follows *t*-distribution with $(n_1 + n_2 - 2)$ degrees of freedom, where

$$S^2 = \frac{1}{n_1 + n_2 - 2}\left[\sum (x_i - \bar{x}^2) + \sum (y_i - \bar{y}^2)\right]$$

$$= \frac{1}{7 + 7 - 2}[56.54 + 78.5488]$$

$$= \frac{1}{12}[135.0888]$$

$$= 11.2574$$

$$T = \frac{2.1 - 3.88}{\sqrt{11.2574\left(\frac{1}{7} + \frac{1}{7}\right)}}$$

$$= \frac{-1.78}{\sqrt{3.2164}}$$

$$= \frac{-1.78}{1.7934}$$

$$= -0.9925$$

Tabulated value of t for 0.05 level of signification and $(n_1 + n_2 - 2) = 12$ degrees of freedom

$t_\alpha = t_{0.05}(12) = 1.782$

Conclusion: Since $|t_{cal}| < t_\alpha$ at 0.05 level of signification, we accept null hypothesis. Hence, the two brands do not differ in aflatoxin content.

(c) t-statistic with 12 degrees of freedom = 0.9925

F-statistic with (1, 12) degrees of freedom = 0.9914

$t^2(\gamma) = 0.985$

$\simeq 0.99$

$\therefore t^2(12) \simeq F(1, 12)$

14.3 THE ANALYSIS FROM DECOMPOSITION OF THE INDIVIDUAL OBSERVATIONS

Let the grand mean $\bar{y} = \frac{\bar{T}}{N} = \frac{\sum_{i=1}^{K}\sum_{j=1}^{n_i} y_{ij}}{N}$

Each observation y_{ij} will be decomposed as $y_{ij} = \bar{y} + (\bar{y}_i - \bar{y}) + (y_{ij} - \bar{y}_i)$ y_{ij} is the observation to be decomposed $\bar{y}$ is the grand mean $(\bar{y}_i - \bar{y})$ is called deviation due to treatment and $(y_{ij} - \bar{y}_i)$ is the error.

Example:

						Total
Method *A*	70	74	64	68	69	345
Method *B*	88	78	84	82	83	415
Method *C*	69	74	73	76	73	365

Observation:

$$\underset{y_{ij}}{\begin{bmatrix} 70 & 74 & 64 & 68 & 69 \\ 88 & 78 & 84 & 82 & 83 \\ 69 & 74 & 73 & 76 & 73 \end{bmatrix}} = \underset{\text{Grand mean}}{\begin{bmatrix} 75 & 75 & 75 & 75 & 75 \\ 75 & 75 & 75 & 75 & 75 \\ 75 & 75 & 75 & 75 & 75 \end{bmatrix}} +$$

$$\underset{\text{Treatment effects } \bar{y}_i - \bar{y}}{\begin{bmatrix} 270 & 270 & 270 & 270 & 270 \\ 340 & 340 & 340 & 340 & 340 \\ 290 & 290 & 290 & 290 & 290 \end{bmatrix}} + \underset{\text{Error } y_{ij} - \bar{y}_i}{\begin{bmatrix} -275 & -271 & -281 & -277 & -276 \\ -327 & -337 & -331 & -333 & -332 \\ -296 & -291 & -292 & -289 & -292 \end{bmatrix}}$$

Hence for example $y_{ij} = 70$

$$\begin{aligned} 70 &= 75 + (69 - 75) + (70 - 69) \\ &= 75 + (-6) + 1 \\ &= 70 \end{aligned}$$

Worked Out Examples

EXAMPLE 14.3

Several different aluminum alloys are under consideration for use in heavy duty circuit wiring applications. Among the desired properties low tested electrical resistance and specimens of each wire are tested by applying a fixed voltage to a given length of wire and measuring the current passing through the wire. Given the following results, would you conclude that these alloys differ in resistance at 0.01 level of significance.

	Alloy I	Alloy II	Alloy III	Alloy IV
Current (amperes)	1.085	1.051	0.985	1.101
	1.016	0.993	1.001	1.015
	1.009	1.022	0.990	
	1.034		0.998	
			1.011	

Solution:

(i) **Null hypothesis:** $\mu_1 = \mu_2 = \mu_3 = \mu_4$

(ii) **Alternative hypothesis:** The μ's are not all equal

(iii) **Level of significance:** $\alpha = 0.01$

(iv) **Criterion:** Tabulated value of F for $(K - 1, N - K) = (4 - 1, 14 - 4) = (3, 10)$ degrees of freedom = 6.55

Reject Null hypothesis if F > 6.55

(v) **Calculation:** The total of the observations in each of 4 samples where sample 1 contains 4 observations

$T_1 = 4.144, n_1 = 4$

$T_2 = 3.066, n_2 = 3$

$T_3 = 4.975, n_3 = 5$

$T_4 = 2.116, n_4 = 2$

Grand Total $\bar{T} = T_1 + T_2 + T_3 + T_4$

$= 14.301$

Total sample size $= n_1 + n_2 + n_3 + n_4$

$N = 4 + 3 + 5 + 2$

$= 14$

Correction term for the mean $C = \dfrac{\bar{T}^2}{N}$

$= \dfrac{14.301^2}{14}$

$= \dfrac{204.518}{14}$

$= 14.6085$

$$SST = \sum_{i=1}^{K} \sum_{j=1}^{n_i} y_{ij}^{\ 2} - C$$

$$= \sum_{i=1}^{4} \sum_{j=1}^{n_i} y_{ij}^{\ 2} C - 14.6085$$

$$= [1.085^2 + 1.016^2 + ... + 1.015^2] - 14.6085$$

$$= 15.6048 - 14.6085$$

$$= 0.9963$$

$$SS(Tr) = \sum_{i=1}^{K} \frac{T_i^{\ 2}}{n_i} - C = \sum_{i=1}^{4} \frac{T_i^{\ 2}}{n_i} - C$$

$$= \left[\frac{4.144^2}{4} + \frac{3.066^2}{3} + \frac{4.975^2}{5} + \frac{2.116^2}{2} \right] - 14.6085$$

$$= 14.6154 - 14.6085 = 0.0069$$

$$SSE = SST - SS(Tr)$$

$$= 0.9963 - 0.0069$$

$$= 0.9894$$

The table of ANOVA is as follows:

Source of variation	Sum of squares	Degrees of freedom	Mean square	F
Alloys	$SS(Tr) = 0.0069$	$K - 1 = 4 - 1 = 3$	$\dfrac{0.0069}{3} = 0.0023$	$F = \dfrac{0.0023}{0.0989}$ $= 0.023$
Error	$SSE = 0.9894$	$N - K = 14 - 4$ $= 10$	$\dfrac{0.9894}{10} = 0.0989$	

(vi) **Decision:** The calculated value of $F = 0.023$. The tabulated value of F for (3, 10) degrees of freedom at 0.05 level of significance

$$F_{\alpha}(3, 10) = 6.55$$

Since the calculated F is not greater than $F_{\alpha}(3, 10)$ at 0.05 level of significance, we accept null hypothesis. Hence, the four alloys do not differ in resistance at 0.01 level of significance.

EXAMPLE 14.4

Prepare an ANOVA table from the following data:

Sample I	77	70	63	84	95	81	88	101
Sample II	109	106	137	79	134	78	126	98
Sample III	46	70	71	65	61	40	47	73
F	17.11							

Solution: The total of the observations in each of the samples where

T_1 = Total of Sample I observations

= 659, $n_1 = 8$

T_2 = Total of Sample II observations

= 867, $n_2 = 8$

T_3 = Total of Sample III observations = 473, $n_3 = 8$

Grand Total $\bar{T} = T_1 + T_2 + T_3 = 1999$

Total sample size = $n_1 + n_2 + n_3$

$N = 8 + 8 + 8 = 24$

Correction term for the mean $C = \dfrac{\bar{T}^2}{N} = 166500.0417$

$$SST = \sum_{i=1}^{3}\sum_{j=1}^{n_i} y_{ij}^{\ 2} - C = 182173 - 166500.0417$$

$$= 15672.9583$$

$$SS(Tr) = \sum_{i=1}^{3} \frac{T_i^2}{n_i} - C = 9712.333$$

$$SSE = SST - SS(Tr) = 5960.6253$$

Table of ANOVA is as follows:

Source of variation	Sum of squares	Degrees of freedom	Mean square	F
Samples	9712.333	3 – 1 = 2	4856.1665	17.1088
Error	5960.6253	24 – 3 = 21	283.8393	

EXAMPLE 14.5

Three training methods were compared to see if they led to greater productivity after training. The productivity measures for individuals trained by different methods that are as follows:

Method I	36	26	31	20	34	25
Method II	40	29	38	32	39	34
Method III	32	18	23	21	33	27

At 0.05 level of significance, do the three training methods lead to different levels of productivity?

Solution:

(i) **Null hypothesis:** $\mu_1 = \mu_2 = \mu_3$

(ii) **Alternative hypothesis:** H_i: μ_i's are not all equal.

(iii) **Level of significance:** $\alpha = 0.05$

(iv) **Criterion:** Reject null hypothesis if $F > F_\alpha$ at α level of significance.

(v) **Calculation:** Calculating the row totals, we get

$T_1 = 172, \quad n_1 = 6$

$T_2 = 212, \quad n_2 = 6$

$T_1 = 154, \quad n_3 = 6$

Grand Total $\bar{T} = 538$

$N = 18$

Correction term for the mean, $C = \dfrac{\bar{T}^2}{N} = 16080.22$

Treatment sum of squares, $SST = \sum_{i=1}^{3}\sum_{i=1}^{n_i} y_{ij}^2 - C$

$= 16836 - 16080.22 = 755.78$

Treatments sum of squares, $SS(Tr) = \sum \dfrac{T_i^2}{n_i} - C$

$$SS(Tr) = \frac{1}{6}\left[172^2 + 212^2 + 154^2\right] - C$$

$= 293.78$

Error sum of squares, $SSE = SST - SS(Tr)$

$= 462$

One-way ANOVA table is as follows:

Source of variation	Sum of squares	Degree of freedom	Mean squares	F
Methods	$SS(Tr) = 293.78$	$K - 1 = 3 - 1 = 2$	$\dfrac{293.78}{2} = 146.89$	$F = \dfrac{146.89}{30.8} = 4.769$
Error	$SSE = 462$	$N - K = 18 - 3 = 15$	$\dfrac{462}{15} = 30.8$	
Totals	$SST = 755.78$	$N - 1 = 17$		

(vi) **Decision:** The calculated value of $F = 4.769$. The tabulated value F for $(K-1, N-K) = (2, 15)$. Degrees of freedom at 0.05 level of significance is

$$F_{0.05}(2,15) = 3.68$$

Since the calculated value is greater than tabulated value of F, we reject null hypothesis at 0.05 level of significance. Hence, the training methods did not lead to different levels of productivity.

EXAMPLE 14.6

Suppose there are three different methods of teaching English that are used on three groups of students. Random samples of size 4 are taken from each group and the marks obtained by the sample students in each group are given. Use ANOVA to find out whether teaching methods had any effects on the student's performance at 0.01 levels

Group *A*	16	17	13	18
Group *B*	15	16	13	17
Group *C*	15	14	13	14

Solution:

(i) **Null hypothesis:** H_0: $\mu_A = \mu_B = \mu_C$

(ii) **Alternative hypothesis:** H_1: μ_i's are all not equal.

(iii) **Level of significance:** $\alpha = 0.01$

(iv) **Criterion:** Reject null hypothesis if $F > F_\alpha$ at α level of significance.

(v) **Calculation:** Calculating the row totals, we get

$T_1 = 64, \quad n_1 = 4$
$T_2 = 60 \quad n_2 = 4$
$T_3 = 56, \quad n_3 = 4$
$\overline{T} = 180, \quad N = 12$

Correction term for the mean $C = \dfrac{\overline{T}^2}{N} = 2700$

$$SST = \sum_{i=1}^{3}\sum_{j=1}^{n_i} y_{ij}^{\,2} - C = 2732 - 2700$$
$$= 32$$
$$SS(Tr) = \sum_{i=1}^{3}\frac{T_i^2}{n_i} - C$$
$$= \frac{1}{4}\left[64^2 + 60^2 + 56^2\right] - C$$
$$= 2708 - 2700$$
$$= 8$$
$$SSE = SST - SS(Tr)$$
$$= 32 - 8 = 24$$

Table of ANOVA is as follows:

Source of variation	Sum of squares	Degree of freedom	Mean squares	F
Groups	$SS(Tr) = 8$	$K - 1 = 3 - 1 = 2$	$\frac{8}{2} = 4$	$F = \frac{4}{2.67} = 1.498$
Error	$SSE = 24$	$N - K = 12 - 3 = 9$	$\frac{24}{9} = 2.67$	

(vi) **Decision:** The calculated value of F = 1.498, the tabulated value of F for (2, 9) degrees of freedom at 0.01 level of significance is $F_{0.01}(2, 9) = 8.09$.

Since the calculated value of F is greater than $F_{0.01}(2, 9)$ at 0.01 level of significance, we accept null hypothesis. Hence, the teaching methods do not have any effects on the student's performance.

EXAMPLE 14.7

The following table gives the words per minute entered in a word processor using three different keyboard positions with the data being obtained after the individuals in each sample had an equal amount of time to familiarize themselves with the keyboard position to be used. Test the null hypothesis that the mean words per minute achieved for the three keyboard positions is not different using 5% level of significance.

Type writer brand

	A_1	A_2	A_3
Number of words per minute	79	74	81
	83	85	65
	62	72	79
	51		55
	77		

Solution:

(i) **Null hypothesis:** H_0: $\mu_1 = \mu_2 = \mu_3$

(ii) **Alternative hypothesis:** H_1: μ_i's are not all equal

(iii) **Level of significance:** $\alpha = 0.05$

(iv) **Criterion:** Reject null hypothesis, if $F > F_\alpha$ at α level of significance.

(v) **Calculation:** Calculating the column totals, we get

$T_1 = 352, \quad n_1 = 5$
$T_2 = 231, \quad n_2 = 3$
$T_1 = 280, \quad n_3 = 4$

Total of all observations, $\overline{T} = 863$

Total of sample sizes, $N = 12$

$$\text{Correction term for the mean } C = \frac{\overline{T}^2}{N} = 62064.083$$

$$\text{Total sum of squares, } SST = \sum_{i=1}^{3}\sum_{j=1}^{n_i} y_{ij}^2 - C$$

$$= 63441 - 62067.083$$

$$= 1376.917$$

$$\text{Treatment sum of squares, } SS(Tr) = \sum \frac{T_i^2}{n_i} - C$$

$$= 103.717$$

$$\text{Error sum of squares, } SSE = SST - SS(Tr)$$

$$= 1273.2$$

One-way AVONA table

Source of variation	Sum of squares	Degree of freedom	Mean squares	F
Type writer brands	$SS(Tr) = 103.717$	$K - 1 = 3 - 1 = 2$	$\frac{103.717}{2} = 51.858$	$F = \frac{51.858}{141.46} = 0.366$
Error	$SSE = 1273.2$	$N - K = 12 - 3 = 9$	$\frac{1273.2}{9} = 141.46$	

(vi) **Decision:** $F_\alpha(K - 1, N - K) = F_{0.05}(2, 9) = 4.26$.

Since $F = 0.366$ is not greater than $F_{0.05}(2, 9)$, we accept null hypothesis at 0.05 level of significance. Hence, there is no significant different between the mean words per minute achieved by three different type writer brands.

EXAMPLE 14.8

An experiment was designed to study the performance of four different detergents for cleaning fuel injectors. The following cleanness readings were obtained with specially designed equipment for 12 tanks of gas distributed over three different models of engines: Apply ANOVA at 0.01 fuel to test whether there are differences in the detergents or in the engines.

	Engine I	Engine II	Engine III	Total
Detergent *A*	45	43	51	139
Detergent *B*	47	46	52	145
Detergent *C*	48	50	55	153
Detergent *D*	42	37	49	128
Totals	182	176	207	565

Solution:

(i) **Null hypothesis:** $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 0$

$\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$

(ii) **Alternative hypothesis:** The α' s and β' s are not all equal to zero.

(iii) **Level of significance:** $\alpha = 0.01$

(iv) **Criterion:** For treatments, reject null hypothesis if $F > F_{0.01}$ $(a - 1, (a - 1)(b - 1))$ (i.e.,) $F > 9.78$.

For blocks, reject null hypothesis if $F > F_{0.01}$ $(b - 1, (a - 1)(b - 1))$ (i.e.,) $F > 10.92$.

(v) **Calculation:** Calculating the row totals, we get

$$T_{1.} = 139,\ T_{2.} = 145,\ T_{3.} = 153,\ T_{4.} = 128$$

$$T_{.1} = 182,\ T_{.2} = 176,\ T_{.3} = 207,$$

$$a = 4,\ b = 3$$

Correction term, $C = \dfrac{\overline{T}^2}{N} = \dfrac{565^2}{12}$

$$= 26602$$

$$SST = \sum\sum y_{ij}^{\ 2} - C = 26867 - 26602 = 265$$

$$SS(Tr) = \frac{1}{3}\left[139^2 + 145^2 + 153^3 + 128^2\right] - 26602$$

$$= 111$$

$$SS(Bl) = \frac{1}{4}\left[182^2 + 176^2 + 207^2\right] - 26602$$

$$= 135$$

$$SSE = SST - SS(Tr) - SS(Bl) = 19$$

Table of ANOVA is as follows:

Source of variation	Sum of squares	Degrees of freedom	Mean square	F
Detergents	111	3	37.0	11.6
Engines	135	2	67.5	21.1
Error	19	6	3.2	

(vi) **Decision:** Since $F_{(Tr)} = 11.6 > F_{0.01}(3, 6)$. Hence, we reject null hypothesis. There are differences in the effectiveness of four detergents.

Since $F_{(Bl)} = 21.1$ exceeds $F_{0.01}(2, 6)$ we reject null hypothesis. There are significant differences among the results obtained for three engines.

Work Book Exercises

1. What assumptions are to be made while using the technique of ANOVA?
2. Explain briefly the ANOVA technique.
3. The following data relates to the production of three varieties of wheat in kilograms shown in 12 plots:

A	14	16	18		
B	14	13	15	22	
C	18	16	19	19	20

   Is there any significant difference in the production of three varieties?
4. Explain the one-way classification technique of ANOVA.
5. Three different methods of teaching statistics are used on their groups of students. Random samples of size 5 are taken from each group and the results are given in the following table. The grades are on a 10-point scale.

Group *A*	Group *B*	Group *C*
7	3	4
6	6	7
7	5	7
7	4	4
8	7	8

   Determine on the basis of the above data whether there is a difference in the teaching methods.
6. A Research company has designed three different systems to clean up oil spills. The following table contains the results measured by how much surface area (in square metres) is cleared in 1 hour. The data were found by testing each method in several trials. Are the three systems equally effective? Use 0.05 level of significance.

System *A*	55	60	63	56	59	55
System *B*	57	53	64	49	62	
System *C*	66	52	61	57		

[**Ans.:** $F = 0.17$]

7. A study compared the number of hours of relief provided by five different brands of antacid administered to 25 different people, each with stomach acid considered strong. The results are as follows:

Brand	*A*	*B*	*C*	*D*	*E*
	4.4	5.8	4.8	2.9	4.6
	4.6	5.2	5.9	2.7	4.3
	4.5	4.9	4.9	2.9	3.8
	4.1	4.7	4.6	3.9	5.2
	3.8	4.6	4.3	4.3	4.4

(i) Compute the mean number of hours of relief for each brand and determine the grand mean.

(ii) Calculate the *F*-ratio. At 0.05 level of significance do the brands produce significantly different amounts of relief to people with strong stomach acid?

[**Ans.:** (i) 4.28, 5.04, 4.90, 3.34, 4.46 and 4.404
(ii) $F = 1.47$. Reject H_0]

8. In Bangalore, a fast food centre feels it is gaining a bad reputation because it takes too long to serve the customers. Because, the centre has four restaurants in this city, it is concerned with whether all four restaurants have the same average service time. One of the owners of the fast food chain has decided to visit each of the stores and monitor the service time for five randomly selected customers. At his forenoon time visits, he records the following service times in minutes:

Restaurant 1	3	4	5.5	3.5	4
Restaurant 2	3	3.5	4.5	4	5.5
Restaurant 3	2	3.5	5	6.5	6
Restaurant 4	3	4	5.5	2.5	3

(i) Using 0.05 level of significance, do all the restaurants have the same mean service time?

(ii) Based on his results, should the owner make any policy recommendations to any of the restaurant managers?

[**Ans.:** (i) $F = 0.51$. Accept H_0
(ii) Because no restaurant is significantly worse than others, any recommendations would have to be made to all the managers]

9. A study compared the effects of four 1-month point of purchase promotions on sales. The unit sales for five stores using all four promotions in different months are as follows:

Free sample	78	87	81	89	85
One-pack gift	94	91	87	90	88
Rupees off	73	78	69	83	76
Refund by mail	79	83	78	69	81

Calculate F-ratio at 0.01 level of significance. Do the promotions produce different effects on sales?
[**Ans.:** $F = 9.69$]

10. The following table gives the data on the performance of three different detergents at three different water temperatures. The performance was obtained on the whiteness readings based on specially designed equipment for nine loads of washing.

	Detergent *A*	Detergent *B*	Detergent *C*
Cold water	45	43	55
Warm water	37	40	56
Hot water	42	44	46

Perform a two–way ANOVA using the level of significance $\alpha = 0.05$.

14.4 TWO-WAY CLASSIFICATION

Next let us consider a two-way ANOVA so that we have two controllable factors.

The experimental error as observed in previous section can be reduced by dividing the observations in each classification into blocks. This is accomplished when known sources of variability are fixed in each block but vary from block to block.

Let y_{ij} denote the observation of the i^{th} treatment and j^{th} block.

Two-way ANOVA table

	Blocks				
	B_1	$B_2 \cdots$	$B_j \cdots$	B_b	Means
Treatment 1	y_{11}	$y_{12} \cdots$	$y_{1j} \cdots$	y_{1b}	$\bar{y}_{1\bullet}$
Treatment 2	y_{21}	$y_{22} \cdots$	$y_{2j} \cdots$	y_{2b}	$\bar{y}_{2\bullet}$
⋮	⋮	⋮	⋮	⋮	
Treatment i	y_{i1}	$y_{i2} \cdots$	$y_{ij} \cdots$	y_{ib}	$\bar{y}_{i\bullet}$
⋮					
Treatment a	y_{a1}	$y_{a2} \cdots$	$y_{aj} \cdots$	y_{ab}	$\bar{y}_{a\bullet}$
Means	$\bar{y}_{\bullet 1}$	$\bar{y}_{\bullet 2} \cdots$	$\bar{y}_{\bullet j} \cdots$	$\bar{y}_{\bullet b}$	$\bar{y}$

This arrangement is also called randomized block design.

Here, $\bar{y}_{i.}$ is the total of i^{th} row and $\bar{y}_{.j}$ is the total of j^{th} column and $\bar{y}$ is the grand total of all observations.

Theorem 14.2

$$\sum_{i=1}^{a}\sum_{j=1}^{b}(y_{ij}-\bar{y})^2 = \sum\sum(y_{ij}-\bar{y}_{i.}-\bar{y}_{.j}+\bar{y})^2 + b\sum_{i=1}^{a}(\bar{y}_{i.}-\bar{y})^2 + a\sum_{j=1}^{b}(\bar{y}_{.j}-\bar{y})^2$$

Proof:

Consider,

$$y_{ij}-\bar{y} = (y_{ij}-\bar{y}_{i.}-\bar{y}_{.j}+\bar{y})+(\bar{y}_{i.}-\bar{y})+(\bar{y}_{.j}-\bar{y})$$

The LHS of the above theorem is called total sum of squares (*SST*). The first term on RHS is called Error Sum of Squares (*SSE*), second term is Treatment Sum of Squares (*SS*(*Tr*)), and the third term is Block Sum of Squares *SS*(*Bl*).

(i.e.,) $SST = SSE + SS(Tr) + SS(Bl)$

(i.e.,) Error sum of squares is $SSE = SST - SS(Tr) - SS(Bl)$

For easy calculations, the sum of squares for two-way ANOVA is given as follows:

$$SST = \sum_{i=1}^{a}\sum_{j=1}^{b} y_{ij}^{\ 2} - C$$

$$SS(Tr) = \frac{\sum_{i=1}^{a} T_{i.}^{\ 2}}{b} - C, \quad T_{i.} \text{ is sum of '}b\text{' observations}$$

$$SS(Bl) = \frac{\sum_{j=1}^{b} T_{.j}^{\ 2}}{a} - C, \quad T_{.j} \text{ is sum of '}a\text{' observations}$$

where *C* is the correction term given by

$$C = \frac{\bar{T}^2}{ab}$$

The two-way ANOVA can be listed in a table as follows:

Sources of variation	Sum of squares	Degrees of freedom	Mean square	*F*-ratio
Treatments	$SS(Tr)$	$a-1$	$MS(Tr) = \frac{SS(Tr)}{a-1}$	$F_{Tr} = \frac{MS(Tr)}{MSE}$
Blocks	$SS(Bl)$	$b-1$	$MS(Bl) = \frac{SS(Bl)}{b-1}$	$F_{Bl} = \frac{MS(Bl)}{MSE}$
Error	SSE	$(a-1)(b-1)$	$MSE = \frac{SSE}{(a-1)(b-1)}$	

We reject null hypothesis. α_i are equal to zero if F_{Tr} exceeds F_α with $(a-1)$ and $(a-1)$ $(b-1)$ degrees of freedom. The null hypothesis that β_j are all equal to zero is rejected if F_{Bl} exceeds F_α with $(b-1)$ and $(a-1)$ $(b-1)$ degrees of freedom.

Caution:

- Any observation corresponding to each treatment within each block is

$$y_{ij} = \mu + + \alpha_i + \beta_j + \in_{ij}$$

where μ = grand mean, α_i = effect of i^{th} treatment, β_j = effect of j^{th} block and $\in_{ij}$ are independent, normally distributed random variables having zero means and common variance σ^2.

Here, in two-way classification, we are interested to test the significance of the differences of $\bar{y}_{i.}$, that is, to test the null hypothesis.

$$\alpha_1 = \alpha_2 = \ldots = \alpha_a = 0$$

In addition, we may test whether the blocking has been effective, that is, to test null hypothesis,

$$\beta_1 = \beta_2 = ... = 0$$

In both these cases alternative hypothesis is at least one of the effects is different from zero.

- Each observation y_{ij} is decomposed as

$$y_{ij} = \bar{y} + (\bar{y}_{i.} - \bar{y}) + (\bar{y}_{.j} - \bar{y}) + (y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y})$$

Observation = grand mean + deviation due to treatment + deviation due to blocks + error

Worked Out Examples

EXAMPLE 14.9

Four different forms of a standardized reading achievement test were given to each of 5 students, and the following scores are obtained:

	Student I	Student II	Student III	Student IV	Student V
Form *A*	75	73	59	69	84
Form *B*	83	72	56	70	92
Form *C*	86	61	53	72	88
Form *D*	73	67	62	79	95

Perform a two-way ANOVA to test at level of significance $\alpha = 0.01$ whether it is reasonable to treat the four forms as equivalent.

Solution: Looking on the forms as treatments and students as blocks, we solve the problem as follows:

(i) **Null hypothesis:** $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 0$

$$\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$$

(ii) **Alternative hypothesis:** The α's and β's are not all equal to zero

(iii) **Level of significance:** $\alpha = 0.01$

(iv) **Criterion:** For treatments, reject null hypothesis, if $F > F_{\alpha}(a-1, (a-1)(b-1))$ degrees of freedom. For blocks, reject null hypothesis if $F > F_{\alpha}(b-1, (a-1)(b-1))$ degrees of freedom.

(i.e.,) Reject H_0 if $F > F_{0.01}(4-1, (4-1)(5-1))$

(i.e.,) $F > F_{0.01}(3, 12)$

$F > 5.95$ for treatments

Reject H_0 if $F > F_{0.01}(5-1, (4-1)(5-1))$

(i.e.,) $F > F_{0.01}(4, 12)$

(i.e.,) $F > 5.41$ for blocks

(v) **Calculation:** Calculating the row totals, we get

$$T_{1\cdot} = 360,\ T_{2\cdot} = 373,\ T_{3\cdot} = 360,\ T_{4\cdot} = 376 \text{ where } a = 4 \text{ and } b = 5$$

Calculating the column totals, we get

$$T_{\cdot 1} = 317,\ T_{\cdot 2} = 273,\ T_{\cdot 3} = 230,\ T_{\cdot 4} = 290,\ T_{\cdot 5} = 359$$

Total sum of squares, $SST = \sum_{i=1}^{a}\sum_{j=1}^{b} y_{ij}^{\ 2} - C$

$$= \sum_{i=1}^{4}\sum_{j=1}^{5} y_{ij}^{\ 2} - C$$

Correction term $C = \dfrac{\bar{T}^2}{ab}$

$\bar{T}$ = grand total = 1469

$$\therefore C = \frac{1469^2}{4(5)} = 107898.05$$

$$SST = [75^2 + 83^2 + \ldots + 95^2] - 107898.05$$
$$= 110607 - 107898.05$$
$$= 2708.95$$

Treatment sum of squares, $SS(Tr) = \dfrac{\sum_{i=1}^{a} T_{i\cdot}^{\ 2}}{b} - C$

$$= \frac{1}{5}\left[360^2 + 373^2 + 360^2 + 376^2\right] - C$$

$$SS(Tr) = \frac{1}{5}[539705] - C$$

$$= 42.95$$

Block sum of squares, $SS(Bl) = \dfrac{\sum_{i=1}^{b} T_{\cdot j}^{2}}{a} - C$

$$= \frac{1}{4}[217^2 + 273^2 + 230^2 + 290^2 + 359^2] - C$$

$$= \frac{1}{4}[440899] - C$$
$$= 2326.7$$

Error sum of squares, $SSE = SST - SS(Tr) - SS(Bl)$
$$SSE = 2708.95 - 42.95 - 2326.7$$
$$= 339.3$$

The two-way ANOVA table is as follows:

Source of variation	Sum of squares	Degrees of freedom	Mean square	
Forms	$SS(Tr) = 42.95$	$a - 1 = 4 - 1 = 3$	$MS(Tr) = \frac{42.95}{3} = 14.3167$	$F_{Tr} = \frac{14.3167}{28.275} = 0.506$
Students	$SS(Bl) = 2326.7$	$B - 1 = 5 - 1 = 4$	$MS(Bl) = \frac{2326.7}{4} = 581.675$	$F_{Bl} = \frac{581.675}{28.275} = 9.9815$
Error	$SSE = 339.3$	$(a - 1)(b - 1) = 3(4) = 12$	$MSE = \frac{339.3}{12} = 28.275$	
Total	$SST = 2708.95$	$ab - 1 = 20 - 1 = 19$		

(vi) **Decision:** Since $F_{Tr} = 0.506$ is not greater than $F_{\alpha}(a - 1, (a - 1)(b - 1)) = 5.95$, we accept null hypothesis, that is, there are no differences in the forms of standardized reading achieved test.

In addition, since $F_{Bl} = 9.9815$ is greater than $F_{\alpha}(b - 1, (a - 1)(b - 1)) = 5.41$, we reject null hypothesis. The differences among the 4 forms obtained by 5 students are significant. There is an effect due to the students. So blocking is important.

Caution:

If the previous problem was looked upon as a one-way classification, the test for differences in the forms would not give significant results.

EXAMPLE 14.10

The following table gives the data on the performance of three different detergents at three different water temperatures. The performance was obtained on the whiteness readings based on specially designed equipment for nine loads of washing:

Washing	Detergent *A*	Detergent *B*	Detergent *C*
Cold water	45	43	55
Warm water	37	40	56
Hot water	42	44	46

Perform a two-way analysis of variance using the level of significance $\alpha = 0.05$.

Solution: Let three different water temperatures be treatments and the three detergents be blocks.

(i) **Null hypothesis:** $\alpha_1 = \alpha_2 = \alpha_3 = 0$

$$\beta_1 = \beta_2 = \beta_3 = 0$$

(ii) **Alternative hypothesis:** The α_i's and β_i's are not all zero.

(iii) **Level of significance:** $\alpha = 0.01$

(iv) **Criterion:** For treatments, reject null hypothesis if $F > F_{\alpha}(a-1, (a-1)(b-1))$ degrees of freedom.

(i.e.,) $F > F_{0.01}(2, 2(2))$ $F > F_{0.01}(2, 4)$

If $F > 18$, reject $H_{0.}$

For block, reject H_0 if $F > F\alpha(b-1, \phi - 1)(b-1))$

$F > F_{0.01}(2, 4)$

$F > 18$, reject H_0.

(v) **Calculation:** Calculating the row totals, we get

$T_{1.} = 143$, $T_{2.} = 133$, $T_{3.} = 132$ where $a = 0$, $b = 3$

Calculating the column totals, we get $T_{.1} = 124$, $T_{.2} = 127$, $T_{.3} = 157$

Total sum of squares, $SST = \sum\sum y_{ij}^{\ 2} - C$

$\bar{T}$ = Grand total = 408

Correction term, $C = \dfrac{\bar{T}^2}{ab} = \dfrac{408^2}{9} = 18496$

$SST = 18820 - 18496$

$= 324$

Treatment of Sums, $SS(Tr) = \dfrac{\sum_{i=1}^{a} T_{i\cdot}^2}{b} - C$

$= \dfrac{1}{3}\left[143^2 + 133^2 + 132^2\right] - C = 24.67$

Block sum of squares, $SS(Bl) = \dfrac{\sum_{i=1}^{b} T_{ij}^2}{a} - C$

$= \dfrac{1}{3}\left[124^2 + 127^2 + 157^2\right] - C = 222$

Error sum of squares, $SSE = SST - SS(Tr) - SS(Bl)$

$= 77.33$

The table of two-way ANOVA is as follows:

Source of variation	Sum of squares	Degrees of freedom	Mean Square	F
Water temperatures	$SS(Tr) = 24.67$	$a - 1 = 2$	$MS(Tr) = \frac{24.67}{2} = 12.335$	$F(Tr) = \frac{12.365}{19.33} = 0.638$
Detergents	$SS(Bl) = 222$	$b - 1 = 2$	$MS(Bl) = \frac{222}{2} = 111$	$F(Bl) = \frac{111}{19.33} = 5.742$
Error	$SSE = 77.33$	$(a - 1)(b - 1) = 4$	$MSE = \frac{77.33}{4} = 19.33$	
Total	$SST = 324$	$ab - 1 = 8$		

(i) **Null hypothesis:** $\alpha_1 = \alpha_2 = \alpha_3 = 0$;

$\beta_1 = \beta_2 = \beta_3 = 0$

(ii) **Alternative hypothesis**: The α_i's are not all zero the β_i's are not all zero.

(iii) **Level of significance:** $\alpha = 0.05$

(iv) **Criterion:** For treatments $F_{0.05}(a - 1, (a - 1)(b - 1)) = F_{0.05}(2, 4) = 18$

$\because F_{Tr}$ is not greater than F_{α}, we accept null hypothesis.

(i.e.,) There is no significant difference between the water temperatures.

For blocks, $F_{0.05}(b - 1, (a - 1)(b - 1)) = F_{0.05}(2, 4) = 18$

$\because F_{(Bl)}$ is not greater than F_{α}, we accept null hypothesis, that is, there is no significant difference between the detergents also.

EXAMPLE 14.11

Apply the technique of ANOVA to the following data, relating to yields of four varieties of wheat in three blocks:

Blocks \ Varieties	1	2	3
I	10	9	8
II	7	7	6
III	8	5	4
IV	5	4	4

Solution: Let the blocks given here be treatments and the varieties of be blocks. $a = 4$ and $b = 3$

(i) **Null hypothesis:** $\alpha_1 = \alpha_2 = \alpha_3 = 0$

$$\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$$

(ii) **Alternative hypothesis:** α_i's are not all zero and β_i's are all zero

(iii) **Level of significance:** $\alpha = 0.05$

Calculation:

Calculating the row totals we get,

$$T_{1\cdot} = 27,\ T_{2\cdot} = 20,\ T_{3\cdot} = 17,\ T_{4\cdot} = 13$$

Calculating the column totals we get,

$$T_{\cdot 1} = 30,\ T_{\cdot 2} = 25,\ T_{\cdot 3} = 22$$

$$\text{Grand total } \bar{T} = 77$$

$$\text{Correction term } C = \frac{\bar{T}^2}{ab} = \frac{77^2}{(4)(3)} = 494.088$$

$$\text{Total sum squares } SST = \sum\sum y_{ij}^{\ 2} - C$$

$$= 541 - 494.083 = 46.917$$

$$\text{Treatment sum of squares } SS(Tr) = \frac{1}{b}\sum_{i=1}^{a} T_{i\cdot}^{\ 2}$$

$$= \frac{1}{3}\left[27^2 + 20^2 + 17^2 + 13^2\right] - C$$

$$= 37.917$$

$$\text{Block sum of squares } SS(Bl) = \frac{1}{a}\sum_{i=1}^{b} T_{\cdot i}^{2} - C$$

$$= \frac{1}{4}[30^2 + 25^2 + 22^2] - C$$

$$= 8.167$$

$$\text{Error sum of squares } SSE = SST - SS(Tr) - SS(Bl)$$

$$= 3.833$$

The two-way ANOVA table is as follows:

Source of variation	Sum of squares	Degrees of freedom	Mean Squares	F
Varieties	$SS(Tr) = 34.917$	$a - 1 = 3$	$\frac{34.917}{3} = 11.639$	$F(Tr) = \frac{11.639}{0.638} = 18.243$

Blocks	$SS(Bl) = 8.167$	$b - 1 = 2$	$\frac{8.167}{2} = 4.0835$	$F(Bl) = \frac{4.0835}{0.638} = 6.4004$
Error	$SSE = 3.833$	$(a - 1)(b - 1) = 6$	$\frac{3.833}{6} = 0.638$	
Total	$SST = 46.917$	$ab - 1 = 11$		

(iv) **Criterion and decision:** For treatments, $F_{0.05}(3, 6) = 4.76$

$F(Tr) = 18.243 > F_{0.05}(3, 6)$. Hence, reject null hypothesis. The varieties of wheat are significantly different.

For blocks, $F_{0.05}(2, 6) = 5.14$

$F(Bl) = 6.4004 > F_{0.05}(2, 6)$. Hence, we reject null hypothesis, that is, the blocks are also significantly different.

EXAMPLE 14.12

A certain company had four salesmen *A*, *B*, *C,* and *D* each of whom was sent for a month to three types of areas country side *K*, outskirts of a city *O*, and shopping centre of a city *S*. The sales in hundreds of rupees per month are as follows:

Distributors / Salesmen	*A*	*B*	*C*	*D*
K	30	70	30	30
O	80	50	40	70
S	100	60	80	80

Carry out an ANOVA and interpret your results.

Solution: Let the districts represent 3 treatments and salesmen represent 4 blocks. $a = 3$, $b = 4$.

(i) **Null hypothesis:** $\alpha_1 = \alpha_2 = \alpha_3 = 0$; $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$.

(ii) **Alternative hypothesis:** H_1: The α_i's β_i's are not all zero.

(iii) **Level of significance:** $\alpha = 0.05$

(iv) **Calculation:** Calculating the row total, we get

$$T_{1\cdot} = 160,\ T_{2\cdot} = 240,\ T_{3\cdot} = 320$$

Calculating the row total, we get

$$T_{\cdot 1} = 210,\ T_{\cdot 2} = 180,\ T_{\cdot 3} = 150,\ T_{\cdot 4} = 180$$

Grand Total $\overline{T} = 720$

Correction term $C = \dfrac{\bar{T}^2}{ab} = 43200$

Total sum of squares, $SST = \sum\sum y_{ij}^{\ 2} - C$

$= 6200$

Treatment sum of squares, $SS(Tr) = \dfrac{1}{b}\sum_{i=1}^{a} T_{i\bullet}^{\ 2} - C$

$= \dfrac{1}{4}\left[160^2 + 240^2 + 320^2\right] - C$

$= 3200$

Block sum of squares, $SS(Bl) = \dfrac{1}{a}\sum_{i=1}^{b} T_{\bullet i}^{\ 2} - C$

$= \dfrac{1}{3}\left[210^2 + 180^2 + 150^2 + 180^2\right] - C$

$= 600$

Error sum of squares, $SSE = SST - SS\,(Tr) - SS\,(Bl)$

$= 2400$

Table of ANOVA is as follows:

Source of variation	Sum of squares	Degrees of freedom	Mean Squares	F
Districts	$SS(Tr) = 3200$	$a - 1 = 2$	$\dfrac{3200}{2} = 1600$	$F(Tr) = \dfrac{1600}{400} = 4$
Salesmen	$SS(Bl) = 600$	$b - 1 = 3$	$\dfrac{600}{3} = 200$	$F(Bl) = \dfrac{200}{400} = 0.5$
Error	$SSE = 2400$	$(a - 1)\,(b - 1) = 6$	$\dfrac{2400}{6} = 400$	
Total	$SST = 6200$	$ab - 1 = 11$		

(v) **Criterion and decision:** $F_{0.05}(a - 1, (a - 1)(b - 1))$ for treatments, $F_{0.05}(2, 6) = 5.14$
$F(Tr) > F_{0.05}(2, 6)$, accept null hypothesis. The sales at three different areas are not significantly different.

For blocks, $F_{0.05}(b - 1, (a - 1)(b - 1)) = F_{0.05}(3, 6) = 4.76$ $F_{(Bl)} = 0.5 > F_{\alpha}(3, 6)$. We accept null hypothesis. The salesmen are also not significantly different. The sales carried in different areas of the city are almost the same as well as salesmen who carried sales are also not significantly different.

Work Book Exercises

11. Set up ANOVA table for the following yield per hectare for three varieties of wheat each grown on four plots, at 0.05 level of significance:

Plot of land \ Variety of wheat	A_1	A_2	A_3
1	16	15	15
2	17	15	14
3	13	13	13
4	18	17	14

12. Perform a two-way ANOVA on the following data:

Treatment II \ Treatment I	1	2	3
1	30	26	38
2	24	29	28
3	33	24	35
4	36	31	30
5	27	35	33

13. Four experiments determine the moisture contents of samples of a powder, each man taking a sample from each of six consignments. Their assessments are as follows:

Consignments \ Observes	1	2	3	4	5	6
1	9	10	9	10	11	11
2	12	11	9	11	10	10
3	11	10	10	12	11	10
4	12	13	11	14	12	10

Analyse the data and discuss if there is any significant difference between consignments or observers.

[**Ans.:** $F_{Tr} = 3.70$
$F_{Bl} = 3.0$]

14. Carry out ANOVA for the following data. A company appoints 4 salesmen *A*, *B*, *C,* and *D* and observes their sales in the seasons—Summer, Winter, and Monsoon. The figures give the sales in lakhs.

Season / Salesmen	*A*	*B*	*C*	*D*
Summer	36	36	21	35
Winter	28	29	31	32
Monsoon	26	28	29	29

14.5 COMPLETELY RANDOMIZED DESIGN (CRD)

The *F*-tests used so far tell us if the differences among the several means are significant, but do not tell us whether a given mean differs from other mean.

If an experiment is confronted with *K* means it may at first test for the significant differences among all possible pairs. This requires large number of tests to be performed. To overcome this, there is an approach to multiple comparisons where it is to test all of the differences using 'Duncan multiple range test'. This test compares the range of any set of *p* means with an appropriate 'Least significant range R_p' given by

$$R_p = S_{\bar{y}} . r_p$$

where $S_{\bar{y}}$ is an estimate of $\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}}$ which can be obtained using,

$$S_{\bar{y}} = \sqrt{\frac{MSE}{n}}$$

where MSE–Error Mean Square and *n* is the common sample size of ANOVA.

r_p depends on the desired level of significance and the number of degrees of freedom corresponding to MSE which is tabulated in Appendix *E*.

r_p values are tabulated for $\alpha = 0.05$ and $\alpha = 0.01$ levels of significance for $p = 2, 3 \ldots 10$ and for various degrees of freedom from 1 to 120.

Applying Duncan multiple range test to the Example (14.8)

First arrange the 4 means of detergents in increasing order:

Detergents	*D*	*A*	*B*	*C*
Mean	$\frac{128}{3} = 42.67$	$\frac{139}{3} = 46.3$	$\frac{145}{3} = 48.3$	$\frac{153}{3} = 51.0$

Next using the error mean square *SSE* = 3.2 from ANOVA and common sample size *n* = 12 we compute $S_{\bar{y}} = \sqrt{\frac{3.2}{12}} = 0.267$

Using r_p table values for $p = 2, 3, 4$

p	2	3	4
r_p	4.32	4.50	4.62

The range of all 4 means = 51 − 42.67

= 8.33

which exceeds $r_4 = 4.62$, the least significant range

Now, leaving the lowest mean, *D*

The range for the remaining 3 detergents = 4.7 which still exceeds r_4 whereas the range of 42.67, 46.3, 48.3.

In other words, among triplets of adjacent means, both sets of differences are significant. So far as pairs of adjacent means are concerned, we find that only the difference between 42.7 and 46.3 is significant.

Hence, we conclude that detergent *D* is significantly inferior to any of the others and detergent *A* is significantly inferior to detergent *C*.

14.6 LATIN SQUARE DESIGN (LSD)

Latin squares are extensively used in agricultural trials in order to eliminate fertility trends in two directions simultaneously. The experimental area is divided into plots arranged in a square in such a manner that there are many plots in each row as there are in each column, this number being equal to the treatments. The plots are then assigned to various treatments such that every treatment occurs only once in every row and column.

Suppose three methods for soldering copper electrical leads and two extraneous sources of variability with

(i) Different operators doing the soldering

(ii) Use of difference soldering fluxes

Suppose three operators and fluxes are to be considered:

	Flux 1	Flux 2	Flux 3
Operator 1	A	B	C
Operator 2	C	A	B
Operator 3	B	C	A

This is called Latin square. Each method is applied once by each operator in conjunction with each flux.

Steps in conducting ANOVA for a Latin Square Design

(i) $$C = \frac{(T\cdots)^2}{r \bullet n^2}$$

$$SS(Tr) = \frac{1}{r \bullet n} \sum_{k=1}^{n} T_k^{\,2} - C$$

$$SSR = \frac{1}{r \bullet n} \sum_{i=1}^{n} T_{i\bullet\bullet}^{2} - C \qquad \text{for rows}$$

$$SSC = \frac{1}{r \bullet n} \sum_{j=1}^{n} T_{\bullet j\bullet}^{2} - C \qquad \text{for columns}$$

$$SS(\text{Rep}) = \frac{1}{n^2} \sum_{i=1}^{r} T_{\bullet\bullet l}^{2} - C \qquad \text{for replicates}$$

$$SST = \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{l=1}^{n} y_{ij(k)l}^{2} - C$$

$$SSE = SST - SS\,(Tr) - SSR - SSC - SS(\text{Rep})$$

where $T_{i\bullet\bullet}$ is total of $r.n$ observations in all the i^{th} rows,
$T_{\bullet j\bullet}$ is total of $r.n$ observations in all the j^{th} columns,
$T_{\bullet\bullet l}$ is the total of n^2 observations in the l^{th} replicate,
$T_{(k)}$ is the total of $r.n$ observations of the K^{th} treatment, and
$T_{\bullet\bullet\bullet}$ is the grand total of all $r.n^2$ observations.

The ANOVA table is as follows:

Source of variation	Sums of squares	Degree of freedom	Mean square	F
Treatments	$SS(Tr)$	$n-1$	$MS(Tr) = \frac{SS(Tr)}{n-1}$	$\frac{MS(Tr)}{MSE}$
Rows	SSR	$n-1$	$MSR = \frac{SSR}{n-1}$	$\frac{MSR}{MSE}$
Columns	SSC	$n-1$	$MSC = \frac{SSC}{n-1}$	$\frac{MSC}{MSE}$
Replicates	$SS(\text{Rep})$	$r-1$	$MS(\text{Rep}) = \frac{SS(\text{Rep})}{r-1}$	$\frac{MS(\text{Rep})}{MSE}$
Error	SSE	$(n-1)(rn+r-3)$	$MSE = \frac{SSE}{(n-1)(rn+r-3)}$	
Totals	SST	rn^2-1		

The model equation for Latin squares is as follows:

Suppose $y_{ij(k)}l$ is the observation of i^{th} row, j^{th} column of l^{th} replicate, and k indicates that it pertains to the k^{th} treatment.

$Y_{ij(k)}l = \mu + \alpha_i + \beta_j + \gamma_k + \rho_l + \in_{ij(k)}{}^{l}$ $i, j, k = 1, 2, \ldots n$ and $l = 1, 2, \ldots r$ such that

$$\sum_{i=1}^{n} \alpha_i = 0, \qquad \sum_{j=1}^{n} \beta_j = 0, \qquad \sum_{k=1}^{n} \gamma_k = 0, \qquad \sum_{l=1}^{n} \rho_l = 0$$

μ is the grand mean,
α_i is the effect of i^{th} row,
β_j is the effects of j^{th} column,
γ_k is the effect k^{th} treatments,
ρ_l is the effect l^{th} replicate,
$\in_{ij(k)l}$ are independent, normally distributed random variables with zero means and common variance σ^2.

Worked Out Examples

EXAMPLE 14.13

Suppose two replicates of the soldering experiment were given. Let the three treatments be the three methods for soldering copper electrical leads. The two extraneous sources of variability may be different operators doing the soldering and the use of different solder fluxes. Suppose three operators and fluxes are considered. The results shown below give the tensile force that is required to separate the soldered leads.

Replicate-I

	Flux 1	Flux 2	Flux 3
Operator 1	A14.0	B16.5	C11.0
Operator 2	C9.5	A17.0	B15.0
Operator 3	B11.0	C12.0	A13.5

Replicate-II

	Flux 1	Flux 2	Flux 3
Operator 1	C10.0	B16.5	A13.0
Operator 2	A12.0	C12.0	B14.0
Operator 3	B13.5	A18.0	C11.5

Analyse this as a Latin square and test at 0.01 level of significance whether there are differences in the methods, the operators, the fluxes or the replicates.

Solution:

(i) **Null hypothesis:** $\alpha_1 = \alpha_2 = \alpha_3 = 0$

$\beta_1 = \beta_2 = \beta_3 = 0$

$\rho_1 = \rho_2 = \rho_3 = 0$

(ii) **Alternative hypothesis:** The α_i's, β_i's, γ_i's, and ρ_i's are not all zero.

(iii) **Level of significance:** $\alpha = 0.01$

(iv) **Calculation:**

$n = 3, r = 2$

$T_{1..}$ = Sum of all observations of all the replicates of the first row

$$= 14.0 + 16.5 + 11.0 + 10.0 + 16.5 + 13.0$$
$$= 81$$
$$T_{2..} = 9.5 + 17 + 15 + 12 + 12 + 14$$
$$= 79.5$$
$$T_{3..} = 11 + 12 + 13.5 + 13.5 + 18 + 11.5$$
$$= 79.5$$
$$T_{\cdot 1 \cdot} = 14 + 9.5 + 11.0 + 10 + 12 + 13.5 = 70$$
$$T_{\cdot 2 \cdot} = 16.5 + 17 + 12 + 16.5 + 12 + 18 = 92$$
$$T_{\cdot 3 \cdot} = 11 + 15 + 13.5 + 13 + 14 + 11.5 = 78$$
$$T_{\cdot\cdot 1} = \text{Total of all observations of Replicate I}$$
$$= 119.5$$
$$T_{\cdot\cdot 2} = \text{Total of all observations of Replicate II}$$
$$= 120.5$$
$$T_{(A)} = \text{Total of all } r \cdot n \text{ observations pertaining to } A^{\text{th}} \text{ treatment}$$
$$= 14 + 17 + 13.5 + 13 + 12 + 18$$
$$= 87.5$$
$$T_{(B)} = 16.5 + 15 + 11 + 16.5 + 14 + 13.5$$
$$= 86.5$$
$$T_{(C)} = 11 + 9.5 + 12.0 + 10 + 12 + 11.5$$
$$= 66$$
$$T\ldots = \text{Total of all } r \cdot n^2 \text{ observations}$$
$$= 240$$

$$\text{Correction term } C = \frac{T_{\ldots}^{\,2}}{r \bullet n^2}$$
$$= \frac{240^2}{2 \bullet 9} = 3200$$

$$\text{Treatment sum of squares, } SS(Tr) = \frac{1}{r \bullet n} \sum_{K=1}^{n} (T_K)^2 - C$$
$$= \frac{1}{2 \bullet 3}\left[T_A^2 + T_B^2 + T_C^2\right] - C$$
$$= \frac{1}{6}\left[87.5^2 + 86.5^2 + 66^2\right] - 3200$$
$$= 49.083$$

$$\text{Row sum of squares, } SSR = \frac{1}{r \bullet n} \sum_{i=1}^{3} T_{i\bullet\bullet}^{\,2} - C$$
$$= \frac{1}{2 \bullet 3}\left[T_{1\bullet\bullet}^{\,2} + T_{2\bullet\bullet}^{\,2} + T_{3\bullet\bullet}^{\,2}\right] - C$$
$$SSR = \frac{1}{6}\left[81^2 + 79.5^2 + 79.5^2\right] - C$$
$$= 0.2$$

$$\text{Column sum of squares, } SSC = \frac{1}{r \bullet n}\sum T_{\cdot j \cdot}^{2} - C$$

$$= \frac{1}{2 \cdot 3}\sum_{j=1}^{3} T_{\cdot j \cdot}^{2} - C$$

$$= \frac{1}{6}\left[T_{\cdot 1 \cdot}^{2} + T_{\cdot 2 \cdot}^{2} + T_{\cdot 3 \cdot}^{2}\right] - C$$

$$= \frac{1}{6}\left[70^2 + 92^2 + 78^2\right] = -C$$

$$= 41.3$$

$$\text{Replicate sum of squares, } SS(\text{Rep}) = \frac{1}{n^2}\sum_{l=1}^{r} T_{\cdot\cdot l}^{2} - C$$

$$= \frac{1}{9}\left[T_{\cdot\cdot 1}{}^{2} + T_{\cdot\cdot 2}{}^{2}\right] - C$$

$$= \frac{1}{9}\left[119.5^2 + 120.5^2\right] - C$$

$$= 0.1$$

$$\text{Total sum of squares, } SST = \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{l=1}^{r} y_{ij\ (K)l}{}^{2} - C$$

$$= [14^2 + 16.5^2 + 11^2 + \ldots\ 11.5^2] - C$$

$$= 3304.5 - 3200 = 104.5$$

$$\text{Error sum of squares, } SSE = SST - SS(Tr) - SSR - SSC - SS(\text{Rep})$$

$$= 104.5 - 49.083 - 0.2\ 41.3 - 0.1$$

$$= 13.8$$

The ANOVA table for the above data is as follows:

Source of variation	Sum of squares	Degree of freedom	Mean square	F
Treatments	$SS(Tr) = 49.1$	$n - 1 = 3 - 1 = 2$	$\frac{49.1}{2} = 24.6$	$F(Tr) = \frac{24.6}{1.38} = 17.8$
Rows	$SSR = 0.2$	$n - 1 = 2$	$\frac{0.2}{2} = 0.1$	$F(R) = \frac{0.1}{1.38} = 0.072$
Columns	$SSC = 41.3$	$n - 1 = 2$	$\frac{41.3}{2} = 20.6$	$F(C) = \frac{20.6}{1.38} = 14.92$
Replicates	$SS(\text{Rep}) = 0.1$	$r - 1 = 2 - 1 = 1$	0.1	$F(\text{Rep}) = \frac{0.1}{1.38} = 0.072$

(*Continued*)

Source of variation	Sum of squares	Degree of freedom	Mean square	F
Error	$SSE = 13.8$	$(n-1) \times (rn + r - 3) = (3-1)(6 + 2 - 3) = 10$	$\frac{13.8}{10} = 1.38$	
Totals	$SST = 104.5$	17		

(v) **Criterion:** For treatments,

$$\text{Tabulated value of } F = F_{0.01}(n-1, (n-1)(nr + r - 3))$$

$$F_{0.01}(2, 10) = 7.56$$

$$\text{For rows, tabulated value of } F = F_{0.01}(2, 10)$$

$$= 7.56$$

$$\text{For columns, tabulated value of } F = F_{0.01}(2, 10)$$

$$= 7.56$$

$$\text{For replicates, tabulated value of } F = F_{0.01}(r-1, (n-1)(rn + r - 3)$$

$$F_{0.01}(1, 10) = 10.04$$

Reject null hypothesis if $F +$ F tabulated value

(vi) **Decision:** For treatments,

$F(Tr) = 17.8$ exceeds $F_{0.01}(2, 10)$ we reject null hypothesis.
For rows, $F(R) = 0.072$ does not exceed, $F_{0.01}(2, 10)$ we accept null hypothesis.
For columns, $F(C) = 14.92$ exceeds $F_{0.01}(2, 10)$ we reject null hypothesis.
For replicates $F_{(rep)} = 0.072$ does not exceed $F_{0.01}(1, 10)$ we accept null hypothesis.

Hence, we conclude that the differences in methods (treatments) and fluxes affect the solder strengths. In addition, the differences in operators or replications do not affect the solder strength of electrical leads.

Remark: Applying Duncan multiple range test for the above problem.

$$\text{Mean of method } A = \frac{T_{(A)}}{6} = 14.58$$

$$\text{Mean of method } B = \frac{T_{(B)}}{6} = 14.41$$

$$\text{Mean of method } C = \frac{T_{(C)}}{6} = 11$$

Thus, method C yields weaker soldiering strength than method A and B.

14.7 RANDOMIZED BLOCK DESIGN (RBD)

Randomized block design is a term that stems from agricultural research in which several variables or treatments are applied to different blocks of land for repetition (replication) of the experimental effects such as yields of different types of soya beans or the quality of different makes of fertilizers.

However, differences in crop yield may be attributed not only to kinds of soya beans but also to differences in quality of the blocks of land. To isolate, the block-effect randomization (which is achieved by assigning treatments at random to plots of each block of land) is employed. The blocks are formed in

such a way that each contains as many plots as there are treatments to be tested and one plot from each block is randomly selected for each treatment.

This RBD is widely used in many types of experiments like in determining the differences in productivity of different makes of machines (treatments). Here, we isolate the possible effects due to differences in efficiency among operators (blocks) by assigning the machines at random to randomly selected operators.

Worked Out Examples

EXAMPLE 14.14

Analyse the following as a Latin square experiment.

Replicate-I

Row \ Column	1	2	3	4
1	*A*(12)	*D*(20)	*C*(16)	*B*(10)
2	*D*(18)	*A*(14)	*B*(11)	*C*(14)
3	*B*(12)	*C*(15)	*D*(19)	*A*(13)
4	*C*(16)	*B*(11)	*A*(15)	*D*(20)

Replicate-II

Row \ Column	1	2	3	4
1	*C*(11)	*B*(12)	*A*(10)	*D*(16)
2	*B*(15)	*D*(13)	*C*(20)	*A*(14)
3	*D*(11)	*A*(16)	*B*(19)	*C*(12)
4	*A*(18)	*C*(15)	*D*(13)	*B*(20)

Solution:

(i) **Null hypothesis:** $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 0$

$\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$

$\gamma_1 = \gamma_2 = \gamma_3 = \gamma_4 = 0$

$\rho_1 = \rho_2 = \rho_3 = \rho_4 = 0$

(ii) **Alternative hypothesis:** The α_i's, β_i's, γ_i's, and ρ_i's are not all zero.

(iii) **Level of significance:** $\alpha = 0.05$

(iv) **Calculation:**

$n = 4, r = 2$

$T_{1..}$ = Sum of all observations of all replicates of the first row

= 12 + 20 +16 +10 +11 + 12 + 10 + 16

= 107

$T_{2..}$ = 18 +14 + 11 +14 +15 + 13 + 20 + 14
= 119
$T_{3..}$ = 12 + 15 + 19 + 13 + 11 + 16 + 19 + 12
= 117
$T_{4..}$ = 16 + 11 + 15 + 20 + 18 + 15 + 13 + 20
= 128
$T_{.1.}$ = Sum of all observations of all replicates of the first column
= 12 + 18 + 12 + 16 + 11 + 15 + 11 + 18 = 113
$T_{.2.}$ = 20 +14 + 15 + 11 + 12 + 13 + 16 + 15 = 116
$T_{.3.}$ = 16 + 11 + 19 + 15 + 10 + 20 + 19 + 13 = 123
$T_{.4.}$ = 10 + 14 + 13 + 20 + 16 + 14 + 12 + 20 = 119
$T_{..1}$ = Total of all observations of Replicate I
= 236
$T_{..2}$ = Total of all observations of Replicate II
= 235
$T_{(A)}$ = Total of all r.n observations pertaining to A^{th} treatment
= 12 + 14 + 13 + 15 + 10 + 14 + 16 + 18
= 112
$T_{(B)}$ = 10 + 11 + 12 + 11 + 12 + 15 + 19 + 20
= 110
$T_{(C)}$ = 16 + 14 + 15 + 16 + 11 + 20 + 12 + 15 = 119
$T_{(D)}$ = 20 + 18 + 19 + 20 + 16 + 13 + 11 + 13 = 130
T = Total of all $r \cdot n^2$ observations
= 236 + 235 = 471

Correction term $C = \dfrac{T...^2}{r.n^2} = \dfrac{471^2}{2.16} = 6932.53$

Treatment sum of squares, $SS(Tr) = \dfrac{1}{r.n}\sum_{K=1}^{n}(T_K)^2 - C$

$= \dfrac{1}{2.4}[T_A{}^2 + T_B{}^2 + T_C{}^2 + T_D{}^2] - C$

$= \dfrac{1}{8}[112^2 + 110^2 + 119^2 + 130^2] - 6932.53$

$= 6963.125 - 6932.53 = 30.595$

Row sum of squares, $SSR = \dfrac{1}{r.n}\sum T_{i..}{}^2 - C$

$= \dfrac{1}{2.4}[T_{1..}{}^2 + T_{2..}{}^2 + T_{3..}{}^2 + T_{4..}{}^2] - C$

$= \dfrac{1}{8}[107^2 + 119^2 + 117^2 + 128^2] - 6932.53$

$= 6960.375 - 6932.53 = 27.845$

Column sum of squares, $SSC = \dfrac{1}{r.n}\sum T.j.^2 - C$

$$= \frac{1}{2.4}[113^2 + 116^2 + 123^2 + 119^2] - 6932.53$$

$$= 6939.375 - 6932.53$$

$$= 6.845$$

Replicate sum of squares, $SS(\text{Rep}) = \frac{1}{n^2}\sum_{l=1}^{r} T_{..l}^{\ 2} - C$

$$= \frac{1}{16}[T_{..1}^{\ 2} + T_{..2}^{\ 2}] - C$$

$$= \frac{1}{16}[236^2 + 235^2] - 6932.53$$

$$= 6932.5625 - 6932.53$$

$$= 0.0325$$

Total sum of squares, $SST = \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\sum_{l=1}^{r} y_{ij\ (kl)}^{\ 2} - C$

$$= [12^2 + 20^2 + 16^2 + \ldots\ldots + 20^2] - 6932.53$$

$$= 7249 - 6932.53$$

$$= 316.47$$

Error sum of squares, SSE = SST − SS(Tr) − SSR − SSC − SS(Rep)
= 316.47 − 30.595 − 27.845 − 6.845 − 0.0325
= 251.1525

The ANOVA table for the above data is as follows:

Source of variation	Sum of squares	Degree of freedom	Mean square	F
Treatments	SS(Tr) = 30.595	$n - 1 = 4 - 1 = 3$	$\frac{30.595}{3} = 10.198$	$F(Tr) = \frac{10.198}{11.959} = 0.853$
Rows	SSR = 27.845	$n - 1 = 3$	$\frac{27.845}{3} = 9.28$	$F(R) = \frac{9.28}{11.959} = 0.776$
Columns	SSC = 6.845	$n - 1 = 3$	$\frac{6.845}{3} = 2.28$	$F(C) = \frac{2.28}{11.959} = 0.191$
Replicates	SS(Rep) = 0.0325	$r - 1 = 1$	0.0325	$F(\text{Rep}) = \frac{0.0325}{11.959} = 0.0027$
Error	SSE = 251.1525	$(n-1) \times (rn + r - 3)$ $= (4-1)(8+2-3)$ $= 21$	$\frac{251.1525}{21} = 11.959$	
Totals	SST = 316.47	16		

(v) **Criterion:** For treatments,

$$\text{Tabulated value of } F = F_{0.05}\,(n-1, (n-1)(nr+r-3))$$

$$F_{0.05}\,(3, 21) = 3.07$$

$$\text{For rows, tabulated value of } F = F_{0.05}\,(3, 21) = 3.07$$

$$\text{For columns, tabulated value of } F = F_{0.05}\,(3, 21) = 3.07$$

$$\text{For replicates, tabulated value of } F = F_{0.05}(r-1, (n-1)(nr+r-3)) = F_{0.05}\,(1, 21) = 4.32$$

We accept null hypothesis as $F > F$ tabulated.

(vi) **Decision:** For treatments,

$F(Tr) = 0.853$ is less than $F_{0.05}\,(3, 21)$ we accept null hypothesis.

For rows, $F(R) = 0.776$ is less than $F_{0.05}\,(3, 21)$, we accept null Hypothesis.

For columns, $F(C) = 0.191$ is less than $F_{0.05}\,(3, 21)$ we accept null Hypothesis.

For replicates, $F(\text{rep}) = 0.0027$ is less than $F_{0.05}\,(1, 21)$ we accept null Hypothesis.

Hence we conclude that there is no effect of rows, columns, treatments and replicates on *A*, *B*, *C*, *D*.

DEFINITIONS AT A GLANCE

Analysis of Variance (ANOVA): A statistical technique used to test the equality of three or more sample means.

One way classification: The ANOVA technique that analysis one variable only.

***SST*:** Total sum of squares.

***SSE*:** Error sum of squares which estimates random error.

***SS*(*TR*):** Treatment of sum of squares.

Two-way classification: The ANOVA technique that involves two factor experiments.

***SS*(*Bl*):** Block sum of squares.

Completely Randomized Design (CRD): A technique where we can know a given mean differs from the other mean.

Latin square Design (LSD): A design which involves Latin square where each method is applied once by each operator.

FORMULAE AT A GLANCE

- For a one-way classification, the overall sample mean $\bar{y} = \dfrac{\bar{T}}{N} = \dfrac{\sum_{i=1}^{K}\sum_{j=1}^{n_i} y_{ij}}{\sum_{i=1}^{K} n_i}$

- Total sum of squares, $SST = \sum_{i=1}^{K} \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2$
- Error sum of squares, $SSE = \sum \sum (\bar{y}_{ij} - \bar{y}_i)^2$
- Treatment sum of squares, $SS(Tr) = \sum_{i=1}^{K} n_i (\bar{y}_i - y)^2$
- For a one-way ANOVA, F-ratio is given by

$$F = \frac{\left(\frac{SS(Tr)}{K-1}\right)}{\left(\frac{SSE}{N-K}\right)} = \frac{MS(Tr)}{MSE} = \frac{\text{Mean square of treatments}}{\text{Mean square of error}}$$

- For a two-way classification,

$$\text{Total sum of squares, } SST = \sum_{i=1}^{a} \sum_{r=1}^{b} y_{ij}^{\ 2} - C$$

$$\text{Treatment sum of squares, } SS(Tr) = \frac{\sum_{i=1}^{a} T_{i\bullet}^{\ 2}}{b} - C$$

$$\text{Block sum of squares, } SS(Bl) = \frac{\sum_{j=1}^{b} T_{\bullet j}^{\ 2}}{a} - C$$

where C is the correction term given by $C = \frac{\bar{T}^2}{ab}$

- For a two-way ANOVA, F-ratio is given by,

$$F_{Tr} = \frac{MS(Tr)}{MSE} = \frac{\left(\frac{SS(Tr)}{a-1}\right)}{\frac{SSE}{(a-1)(b-1)}} \text{ and}$$

$$F_{Bl} = \frac{MS(Bl)}{MSE} = \frac{\left(\frac{SS(Bl)}{b-1}\right)}{\frac{SSE}{(a-1)(b-1)}}$$

- In Duncan multiple range test,

  Least significant range, $Rp = \sqrt{\frac{MSE}{n}} \bullet rp$, where rp is the desired level of significance.
- For C Latin square design, ANOVA has following formulae,

$$\text{Sum of squares treatment, } SS(Tr) = \frac{1}{r \bullet n} \sum_{K=1}^{n} T_K^2 - C$$

$$\text{Sum of squares of rows, } SSR = \frac{1}{r \bullet n} \sum_{i=1}^{n} T_{i\bullet\bullet}^2 - C$$

$$\text{Sum of columns, } SSC = \frac{1}{r \bullet n} \sum_{j=1}^{n} T_{\bullet}^2 j_{\bullet} - C$$

$$\text{Sum of squares of replicates, } SS_{(\text{Rep})} = \frac{1}{n^2} \sum_{i=1}^{r} T_{\bullet\bullet l}^2 - C$$

$$\text{Total sum of squares, } SST = \sum_{i=1}^{n} \quad \sum_{j=1}^{n} \quad \sum_{l=1}^{r} y_{ij(K)l}^2 - C$$

- For LSD, *F*-ratio is given by

$$F_{Tr} = \frac{MS(Tr)}{MSE} = \frac{\left(\frac{SS(Tr)}{n-1}\right)}{\frac{SSE}{(n-1)(rn+r-3)}}$$

$$F_R = \frac{MSR}{MSE} = \frac{\left(\frac{SSR}{n-1}\right)}{\frac{SSE}{(n-1)(rn+r-3)}}$$

$$F_C = \frac{MSC}{MSE} = \frac{\left(\frac{SSC}{n-1}\right)}{(n-1)(rn+r-3)}$$

$$F_{\text{Rep}} = \frac{MS(\text{Rep})}{MSE} = \frac{\left(\frac{SS_{\text{Rep}}}{r-1}\right)}{\frac{SSE}{(n-1)(rn+r-3)}}$$

OBJECTIVE TYPE QUESTIONS

1. In one way ANOVA, the relation between SST, SSE and SS(Tr) is
   (a) SSE = SST + SS(*Tr*) (b) SST = SSE + SS(*Tr*)
   (c) SS(*Tr*) = SSE + SST (d) none

2. If *C* is the correction term for the mean, then SST = _____
   (a) $\sum_i \sum_j y_{ij}^{\ 2} - C$ (b) $\sum_i \sum_j \frac{y_{ij}^{\ 2}}{n} - C$
   (c) $\sum_i \sum_j y_{ij} - C$ (d) none

3. If there are K populations and N denotes the total sample size SSE is the error sum of squares, then error mean square is MSE = _____

(a) $\frac{SSE}{N-1}$ (b) $\frac{SSE}{K-1}$

(c) $\frac{SSE}{N-K}$ (d) none

4. If T_i is the total of n_i observations in the i^{th} sample and C is the correction term, then treatment sum of squares SS(Tr) = _____

(a) $\sum_i \frac{T_i^2}{n_i}$ (b) $\sum \frac{T_i^2}{n_i} - C$

(c) $\sum \frac{T_i}{n_i} - C$ (d) none

5. In a two way ANOVA, the relation between Treatment sum of squares SS(Tr), error sum of squares SSE, block sum of squares SS(Bl) and total sum of squares SST is _____

(a) SST = SSE + SS(Tr) + SS(Bl) (b) SSE = SS(Tr) + SS(Bl)

(c) SS(Tr) = SST + SS(Bl) + SSE (d) none

6. In a two way ANOVA, the mean square error for 'a' treatments and 'b' blocks is given by MSE,

(a) $\frac{SSE}{ab}$ (b) $\frac{SSE}{(a-1)(b-1)}$

(c) $\frac{SSE}{ab-1}$ (d) none

7. In a two way ANOVA, given mean square treatments MS(Tr) mean square blocks and mean square error, F-ratio of blocks is _____

(a) $F_{Bl} = \frac{MS(Tr)}{MS(Bl)}$ (b) $F_{Bl} = \frac{MS(Bl)}{MS(Tr)}$

(c) $F_{Bl} = \frac{MS(Bl)}{MSE}$ (d) none

8. The degrees of freedom for F-test for two way ANOVA is

(a) $(a-1, b-1)$ (b) $(a-1, ab-1)$

(c) $(a-1, (a-1)(b-1))$ (d) none

9. According to Duncan multiple range test, an estimate of $\sigma_{\bar{y}}$, $S_{\bar{y}}$ is given by

(a) $S_{\bar{y}} = \sqrt{\frac{MSE}{n-1}}$ (b) $S_{\bar{y}} = \sqrt{\frac{MSE}{n}}$

(c) $S_{\bar{y}} = \sqrt{\frac{MSE}{n+1}}$ (d) none

10. In a latin square design with m treatments, *n* columns, *r* replicates, the mean square error is given by

(a) $MSE = \frac{SSE}{(n-1)(r-1)}$ (b) $MSE = \frac{SSE}{(n-1)(m-1)(r-1)}$

(c) $MSE = \frac{SSE}{(n-1)(rn+r-3)}$ (d) none

11. A design in which several variables or treatments are applied to different blocks for repetition of the experimentsal effects is called

(a) Latin square design (b) Randomized block design
(c) Completely randomized design (d) none

ANSWERS

1. (b) 2. (a) 3. (c) 4. (b) 5. (a) 6. (b) 7. (c) 8. (c)
9. (b) 10. (c) 11. (b)

15 Random Process

Prerequisites

Before you start reading this unit, you should:

- Be familiar with random variables
- Know expectations, distribution functions, and density functions

Learning Objectives

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

- Distinguish between continuous, discrete, deterministic, and non-deterministic random processes
- Find the cross correlation, Ergodic theorem, and distribution correlation
- Familiar with first order, second order, stationary random process, and mean-ergodic theorem

INTRODUCTION

The concept of random process is based on enlarging the concept of random variable to include time. A random variable is a rule or a function that assigns a real number to every outcome of a random experiment, while a random process is a rule or a function that assigns a time function to every outcome of a random experiment.

Since a random variable X is a function of the possible outcomes s of an experiment, a random process is a function of both s and time. The family of functions denoted by $X(t, s)$ is called a random process.

If in the family of functions, s is fixed, the random process is a function of time only. It is called single time function.

If in the above, t is fixed, then the random process is a function of s only and hence the random process will represent a random variable at time 't'.

If in the above, both t and s are fixed, then $X(t, s)$ is merely a number.

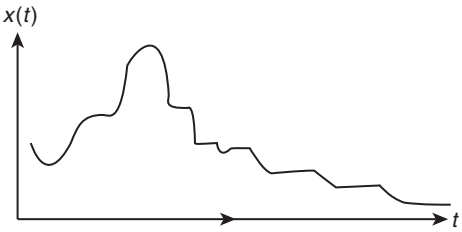
If both t and s are variables, then $X(t, s)$ represents a collection of random variables that are time functions.

15.1 CLASSIFICATION OF RANDOM PROCESSES

Random processes are broadly classified into two categories namely continuous random processes and discrete random processes.

Continuous Random Processes

A continuous random process is one in which the random variable X is continuous and t is also continuous, that is, both t and s are continuous.



Continuous random processes

The practical examples of a continuous random process are thermal noise generated by a realizable network, fluctuations in air temperature, and air pressure, etc.

Discrete Random Process

A random process is said to be discrete random process if s is discrete and t is continuous.

A practical example of a discrete random process is the voltage available at one end of the switch. This is discrete because of random opening and closing of the switch.

The Random processes may also be classified into two, like deterministic or non-deterministic random processes.

Discrete random process

Deterministic Random Processes

A process is called deterministic random process if future values of any sample function can be predicted from past values.

An example is the random process defined by

$$X(t) = A\ \mathrm{Cos}(\omega t + \theta)$$

Here, A, θ, ω may be random variables. Hence, knowledge of the sample function prior to any time instant automatically allows prediction of the sample functions' future values because its form is known.

Non-deterministic Random Process

If the future values of any sample function cannot be predicted exactly from observed past values, the process is called non-deterministic random process.

An example is dissolving of sugar crystals in coffee. It consists of a family of functions that cannot be described in terms of a finite number of parameters. The future sample function cannot be determined from the past sample functions and hence it is non-deterministic random process.

15.2 STATIONARITY

In the earlier section, we have seen that a random process becomes a random variable when time is fixed at some particular value. From the random variable, all the statistical properties such as mean, variance, moments, and so on can be obtained which are related to its density function. If two random variables are obtained from the process at two time instants, all the statistical properties such as means, variances, joint moments, and so on can be obtained related to their joint density function.

Generalizing this, N random variables will possess all the statistical properties related to their N-dimensional joint density function.

Hence, stationary random process can be defined as follows:

A random process is said to be stationary if all its statistical properties do not change with time.

Let us first define distribution and density functions in the analysis of random process.

Distribution and Density Functions

The distribution function associated with the random variable $X_1 = X(t_1)$ for a particular time t_1 is defined as

$$F_X(x_1; t_1) = P\{X(t_1) \le x_1\} \tag{15.1}$$

where x_1 is any real number. This is known as the first order distribution function of the process $X(t)$.

The second order joint distribution function for two random variables $X_1 = X(t_1)$ and $X_2 = X(t_2)$ is the two dimensional extension of the above.

$$F_X(x_1, x_2; t_1, t_2) = P\{X(t_1) \le x_1, X(t_2) \le x_2\} \tag{15.2}$$

Generalizing this we can define for N random variables $X_1 = X(t_1)$, $X_2 = X(t_2)$, … $X_N = X(t_N)$

The N^{th} order joint distribution function is

$$F_X(x_1, x_2, \ldots x_N; t_1, t_2, \ldots t_N) = P\{X(t_1) \le x_1, \ldots X(t_N) \le x_N\} \tag{15.3}$$

The derivatives of these functions give density functions

$$f_X(x_1; t_1) = \frac{d}{dx_1} F_X(x_1; t_1)$$

$$f_X(x_1, x_2; t_1, t_2) = \frac{\partial^2 F_X}{\partial x_1 \partial x_2}(x_1, x_2; t_1, t_2)$$

$$f_X(x_1, x_2, \ldots x_N; t_1, t_2, \ldots t_N) = \frac{\partial^N F_X}{\partial x_1 \ldots \partial x_N}(x_1, x_2, \ldots x_N; t_1, t_2, \ldots t_N)$$

First Order Stationary Process

A random process is called first order stationary process if its first order density function does not change with a shift in time origin.

(i.e.,) $f_X(x_1; t_1) = f_X(x_1; t_1 + \Delta)$ for any t_1 and any real number Δ, if $X(t)$ is a first order stationary process.

Worked Out Example

EXAMPLE 15.1

Prove that a first order stationary random process has a constant mean.

Solution: Consider a random process $X(t)$ at two different instants t_1 and t_2. Mean value of the random variables $X_1 = X(t_1)$ and $X_2 = X(t_2)$ are

$$E(X_1) = E[X(t_1)] = \int_{-\infty}^{\infty} x_1 f_X(x_1; t_1)\, dx_1$$

For X_2,

$$E(X_2) = E[X(t_2)]$$

$$= \int_{-\infty}^{\infty} x_1 f_X(x_1; t_2)\, dx_1$$

Now, let $t_2 = t_1 + \Delta$ in the above equation,

$$E(X_2) = E[X(t_1 + \Delta)]$$

$$= \int_{-\infty}^{\infty} x_1 f_X(x_1; t_1)\, dx_1$$

$$= E[X(t_1)]$$

$$= \text{Constant} \qquad \text{Since } t_1 \text{ and } \Delta \text{ are arbitrary.}$$

15.3 SECOND ORDER STATIONARY PROCESS

A random process is called stationary to order two if its second order density function satisfies $f_X(x_1, x_2; t_1, t_2) = f_X(x_1, x_2; t_1 + \Delta, t_2 + \Delta)$ for all t_1, t_2, and Δ.

A second order stationary process is also first order stationary process.

Before going to the next definition let us define autocorrelation function.

The autocorrelation function of the random process $X(t)$ is given by

$$R_{XX}(t_1, t_2) = E[X(t_1)X(t_2)]$$

which is a function of t_1 and t_2.

Let $t = t_2 - t_1$

The autocorrelation function of a second order stationary process is a function of time differences only and not absolute time.

$$\therefore R_{XX}(t_1, t_1 + t) = E[X(t_1)X(t_1 + t)]$$
$$= R_{XX}(t)$$

15.4 WIDE SENSE STATIONARY PROCESS

A stationary random process $X(t)$ is called wide sense stationary (WSS) random process (also called as weak sense stationary process) if it satisfies the following conditions:

(i) The mean value of the process is a constant

(i.e.,) $$E[X(t)] = \text{Constant} \tag{15.4}$$

(ii) Its autocorrelation function depends only on τ and not on t i.e.,

$$E[X(t)X(t + \tau)] = R_{XX}(\tau) \tag{15.5}$$

Caution:

- It is abbreviated as WSS.
- A process which is stationary to order 2 is always WSS but the converse is not necessarily true.

15.5 CROSS CORRELATION FUNCTION

Let $X(t)$ and $Y(t)$ be two random processes. We say that they are jointly WSS if each satisfies equations (15.4) and (15.5). Then their cross correlation function is given by

$$R_{XY}(t_1, t_2) = E[X(t_1)Y(t_2)]$$

which is a function only of time difference $\tau = t_2 - t_1$ and not absolute time

(i.e.,) $$R_{XY}(t, t + \tau) = E[X(t)Y(t + \tau)]$$
$$= R_{XY}(\tau)$$

15.6 STATISTICAL AVERAGES

Two statistical averages mostly used in the description of a random process are the mean and the autocorrelation function.

The mean of the random process $X(t)$ is the expected value of the random variable X at time t which is given by

$$\tilde{X} = E[X(t)] = \int_{-\infty}^{\infty} X f_X(x, t)\, dx$$

The autocorrelation of the random process $X(t)$ is the expected value of the product $X(t_1)\, X(t_2)$ and is given by

$$R_{X_1 X_2}(t_1, t_2) = E\{X(t_1) X(t_2)\}$$
$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} X_1 X_2 f_{X_1 X_2}(X_1, X_2; t_1, t_2)\, dx_1 dx_2$$

15.7 TIME AVERAGES

The mean and autocorrelation function shown above can also be determined by using time averages. The time averaged mean is defined as

$$<\mu_X> = \underset{T\to\infty}{\text{Lt}} \frac{1}{T} \int_{-\frac{T}{2}}^{\frac{T}{2}} X(t)\, dt$$

and the time averaging auto correlation function is $<R_{XX}(\tau)> = \underset{T\to 0}{\text{Lt}} \frac{1}{T} \int_{-T/2}^{T/2} X(t) X(t+\tau) dt$

where $<\mu_X>$ and $R_{XX}(\tau)$ are random variables.

15.8 STATISTICAL INDEPENDENCE

Two random processes $X(t)$ and $Y(t)$ are said to be statistically independent processes if the group of random variables $X(t_1)$, $X(t_2)$, … $X(t_N)$ is independent of the group of random variables $Y(t'_1)$, $Y(t'_2)$,… $Y(t'_N)$ for any choice of times $t_1, t_2, \ldots t_N$ and $t'_1, t'_2, \ldots t'_M$.

Mathematically this may be written as

$$f_{X,Y}(X_1, X_2, \ldots X_N, Y_1, Y_2, \ldots Y_M; t_1, t_2, \ldots t_N, t'_1, t'_2, \ldots t'_M)$$
$$= f_X(X_1, X_2, \ldots X_N; t_1, t_2, \ldots t_N) f_Y(Y_1, Y_2, \ldots Y_M, t'_1, t'_2, \ldots t'_M)$$

Caution:

Random processes that are not stationary are called non-stationary processes or evolutionary processes.

15.9 ERGODIC RANDOM PROCESS

The ergodicity of a random process is used to estimate the various statistical quantities of a random process $X(t)$, that is, if the mean of a random process is to be estimated, a large number of outcomes $X(t, s_i)$ are taken and their ensemble average is calculated as

$$m(t) = \frac{1}{n} \sum_i X(t, s_i)$$

We normally use ensemble averages (or statistical averages) such as the mean and autocorrelation function for characterizing random processes. To estimate ensemble averages, we have to compute a weighted average over all the membership functions of the random process.

In general, ensemble averages and time averages are not equal except for a very special class of random processes called ergodic processes.

Definition

A random process $\{X(t)\}$ is said to be ergodic, if its ensemble averages are equal to appropriate time averages.

Hence the random process $\{X(t)\}$is said to be mean-ergodic if the random process $X(t)$ has a constant mean $E[X(t)] = \mu$ and if $<\mu_X> = \underset{T\to\infty}{\text{Lt}} \frac{1}{T}\int_{-\frac{T}{2}}^{\frac{T}{2}} X(t')\,dt$ tends to μ.

We can define ergodicity with respect to various statistical parameters such as, mean, variance, correlation, distribution function, and so on. Hence, we have mean-ergodic process, variance-ergodic process, distribution ergodic process, and correlation ergodic process.

15.10 MEAN-ERGODIC THEOREM

If $\{X(t)\}$ is a random process with constant mean μ and if $<\mu_X> = \underset{T\to\infty}{\text{Lt}} \frac{1}{T}\int_{-\frac{T}{2}}^{\frac{T}{2}} X(t)\,dt$ then $\{X(t)\}$ is mean-ergodic provided $\underset{T\to\infty}{\text{Lt}}[\text{Var}<\mu_X>] = 0$

Proof:

$$<\mu_X> = \frac{1}{T}\int_{-\frac{T}{2}}^{\frac{T}{2}} X(t)dt$$

$$E[<\mu_X>] = \frac{1}{T}\int_{-\frac{T}{2}}^{\frac{T}{2}} E(X(t))dt$$

$$= \frac{\mu}{T}\int_{-\frac{T}{2}}^{\frac{T}{2}} dt = \frac{\mu}{T}|t|_{-\frac{T}{2}}^{\frac{T}{2}}$$

$$E[<\mu_X>] = \mu \tag{15.6}$$

By Chebychev's inequality,

$$P\{|<\mu_X> - E(<\mu_X>)| \le E\} \ge 1 - \frac{Var(<\mu_x>)}{E^2} \tag{15.7}$$

Using (15.6) by taking limit $T \to \infty$

$$P\{|\underset{T\to\infty}{\text{Lt}} <\mu_X> - \mu| \le E\} \ge 1 - \frac{\underset{T\to\infty}{\text{Lt}}\ \text{Var}(<\mu_X>)}{E^2} \tag{15.8}$$

$\therefore$ when $\text{Var}(<\mu_X>) = 0$, equation (15.8) becomes

$$P\{|\underset{T\to\infty}{\text{Lt}} <\mu_X> - \mu| \le E\} \ge 1$$

$$\underset{T\to\infty}{\text{Lt}} < \mu_X > = E(< \mu_X >) \quad \text{with probability 1.}$$

Caution:

The above theorem provides a sufficient condition for the mean ergodicity of a random process, that is, to prove mean ergodicity of $\{X(t)\}$, it is enough to prove $\underset{T\to\infty}{\text{Lt}} \text{var}(< \mu_X >) = 0$.

Mean-Ergodic Process

A process $X(t)$ with a constant mean value is called mean-Ergodic if its statistical average $\bar{X}$ equals the time average μ_X of any sample function $X(t)$ with probability 1 for all sample functions.

For any $\varepsilon > 0$, how small it may be then $X(t) = m$ and is of probability 1. Sufficient condition for mean ergodicity is

$$\underset{T\to\infty}{\text{Lt}} V[\overline{X(t)}] = 0$$

$$\mu_X(t) = \frac{1}{T}\int_{-\frac{T}{2}}^{\frac{T}{2}} X(t)\,dt \quad \text{and}$$

$$E[\bar{X}(t)] = E[X(t)]$$

$$\text{Consider,} \quad [X(t)]^2 = \frac{1}{T^2}\int_{-\frac{T}{2}}^{\frac{T}{2}}\int_{-\frac{T}{2}}^{\frac{T}{2}} X(t_1)\cdot X(t_2)\,dt_1 dt_2$$

$$= \frac{1}{4T^2}\int_{-T}^{T}\int_{-T}^{T} X(t_1)X(t_2)\,dt_1 dt_2$$

$$\text{Consider} \quad E[\{X(t)\}^2] = \frac{1}{4T^2}\int_{-T}^{T}\int_{-T}^{T} E[X(t_1)X(t_2)]dt_1 dt_2$$

$$= \frac{1}{4T^2}\int_{-T}^{T}\int_{-T}^{T} R_{XX}(t_1,t_2)dt_1 dt_2$$

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

$$V(\overline{X(t)}) = \frac{1}{4T^2}\int_{-T}^{T}\int_{-T}^{T} R_{XX}(t_1t_2)\,dt_1 dt_2 - \frac{1}{4T^2}\int_{-T}^{T}\int_{-T}^{T} E[X(t_1)]\cdot E[X(t_2)]dt_1 dt_2$$

$$= \frac{1}{4T^2}\int_{-T}^{T}\int_{-T}^{T} [R_{XX}(t_1,t_2) - E[X(t_1)]\cdot E[X(t_2)]]dt_1 dt_2$$

$$= \frac{1}{4T^2}\int_{-T}^{T}\int_{-T}^{T} C_{XX}(t_1,t_2)\,dt_1 dt_2$$

Thus, the condition $\underset{T\to\infty}{\text{Lt}} V[\overline{X(t)}] = 0$ is same as $\underset{T\to\infty}{\text{Lt}} \frac{1}{4T^2}\int_{-T}^{T}\int_{-T}^{T} C_{XX}(t_1,t_2)\,dt_1 dt_2 = 0.$

If $X(t)$ is considered as a WSS process, $C_{XX}(t_1, t_2)$ is a function of $\tau = (t_2 - t_1)$.

$$\therefore\ C_{XX}(t_1, t_2) = C_{XX}(\tau)$$

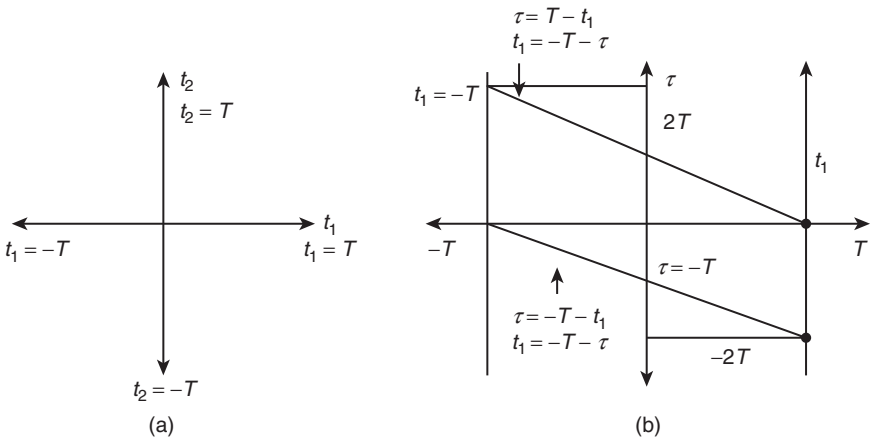
The above double integral is evaluated over a square bounded by $t_1 = -T$ and T and $t_2 = -T$ and T. Let $\tau = t_2 - t_1$ as t_2 varies from $-T$ to T, τ ranges from $(-T - t_1)$ to $(+T - t_1)$ and $d\tau = dt_2$.

$$V[\overline{X(t)}] = \frac{1}{4T^2}\int_{-T}^{T}\int_{(-T-t_1)}^{T-t_1} C_{XX}(\tau)d\tau\,dt.$$

Consider t_1 along x-axis and τ along y-axis. Then $\tau = -T - t$ is a straight line passing through $\tau = -T$ and is equal to $-S$. Similarly, $\tau = T - t_1$ is a straight line passing through $\tau = T$ and is equal to -1.

The variable t_1 ranges from $-T$ to $+T$. Hence, the double integral area is evaluated as follows:

Initially integration is with respect to t_1 and this strip is to be slided horizontally from bottom to top. $(T)_{max} = 2T$ and $(T)_{min} = -2T$ are shown in the following figures:



Since $\tau = T - t_1 \Rightarrow t_1 = T - \tau$

$\tau = -T - t_1 \Rightarrow t_1 = -T - \tau$

$$\text{Var}[\overline{X(t)}] = \frac{1}{4T^2}\int_{T=2T}^{2T}\int_{-T-\tau}^{T-\tau} C_{XX}(\tau)dt_1 d\tau$$

$$V[\overline{X(t)}] = \frac{1}{4T^2}\left[\int_{-2T}^{0}\int_{T-\tau}^{T} C_{XX}(\tau dt_1 d\tau + \int_{0}^{2T}\int_{-T}^{T-\tau} C_{XX}(\tau)dt_1 d\tau\right]$$

$$= \frac{1}{4T^2}\left[\int_{-2T}^{0} C_{XX}(\tau)d\tau \int_{-T-\tau}^{T} dt_1 + \int_{0}^{2T} C_{XX}(\tau)d\tau.\int_{-T}^{T-\tau} dt_1\right]$$

$$= \frac{1}{4T^2}\left[\int_{-2T}^{0} C_{XX}(\tau)d\tau\left[T-(-T-\tau)\right] + \int_{0}^{2T} C_{XX}(\tau)d\tau\left[T-\tau-(-T)\right]\right]$$

$$= \frac{1}{4T^2}\left[\int_{2T}^{0}(2T+\tau)C_{XX}(\tau)d\tau + \int_{0}^{2T}(2T-\tau)C_{XX}(\tau)d\tau\right]$$

$$= \frac{2T}{4T^2}\left[\int_{-2T}^{0}\left(1+\frac{\tau}{2T}\right)C_{XX}(\tau)d\tau + \int_{0}^{2T}\left(1-\frac{\tau}{2T}\right)C_{XX}(\tau)d\tau\right]$$

$$= \frac{1}{2T}\int_{0}^{2T}\left[1-\frac{|T|}{2T}\right]C_{XX}(\tau)d\tau$$

$\{\because C_{XX}(\tau) = C_{XX}(-\tau)\}$

$$\therefore V[\overline{X(t)}] = \frac{1}{T}\int_0^{2T}\left(1-\frac{|T|}{2T}\right)C_{XX}(\tau)d\tau$$

However, sufficient condition is

$$\underset{T\to\infty}{\text{Lt}}\ \frac{1}{T}\int_0^{2T}\left(1-\frac{|\tau|}{2T}\right)C_{XX}(\tau)d\tau = 0$$

It can also be stated as

$$\underset{T\to\infty}{\text{Lt}}\ \frac{1}{2T}\int_{-2T}^{2T} C_{XX}(\tau)\left[1-\frac{|\tau|}{2T}\right]d\tau = 0$$

Since τ is varying between $-2T$ to $2T$

$$\Rightarrow |\tau| \le 2T$$

$$\Rightarrow \frac{1}{2T}\int_{-2T}^{2T} C_{XX}(\tau)\left[1-\frac{|\tau|}{2T}\right]d\tau \le \frac{1}{2T}\int_{-2T}^{2T}|C_{XX}(\tau)|d\tau$$

The above condition will be true only if

$$\underset{T\to\infty}{\text{Lt}}\ \frac{1}{2T}\int_{-2T}^{2T}|C_{XX}(\tau)|d\tau = 0$$

If $\int_{-\infty}^{\infty}|C_{XX}(\tau)|d\tau < \infty$ (i.e.,) finite

Hence the sufficient condition can also be stated as follows:

$$\int_{-\infty}^{\infty}|C_{XX}(\tau)|d\tau < \infty$$

Similarly, a discrete sequence is said to be mean-ergodic if the statistical average of sequence and the time average of samples are equal with a probability of 1.

The statistical average is $E[X(t)]$ and time average is $\underset{N\to\infty}{\text{Lt}}\ \frac{1}{2N+1}\sum_{n=-N}^{N} X(n)$.

In terms of variance of the process, the condition for mean ergodicity is

$$\underset{N\to\infty}{\text{Lt}}\ \frac{1}{2N+1}\sum_{n=-2N}^{2N}\left[1-\frac{|n|}{(2N+1)}\right]C_{XX}(n) = 0$$

15.11 CORRELATION ERGODIC PROCESS

The stationary process $\{X(t)\}$ is said to be correlation ergodic if the process $\{Z(t)\}$ is mean-ergodic where $Z(t) = X(t+\lambda) \times X(t)$, that is, the stationary process $\{X(t)\}$ is correlation ergodic if

$$<\mu_Z> = \underset{T\to\infty}{\text{Lt}}\ \frac{1}{T}\int_{-\frac{T}{2}}^{\frac{T}{2}} X(t+\lambda)X(t)$$

$$= E\{X(t+\lambda)X(t)\}$$

$$= R(\lambda)$$

Distribution Ergodic Process

If $\{X(t)\}$ is stationary process and $\{Y(t)\}$ is another process which is defined as

$$Y(t) = \begin{cases} 1, & \text{if } X(t) \le x \\ 0, & \text{if } X(t) > x \end{cases}$$

then $\{X(t)\}$ is said to be distribution ergodic if $\{Y(t)\}$ is mean-ergodic, that is, the stationary process $\{X(t)\}$ is distribution ergodic, if

$$\begin{aligned} <\mu_Y> &= \underset{T\to\infty}{\text{Lt}} \frac{1}{T} \int_{-\frac{T}{2}}^{\frac{T}{2}} Y(t)dt \\ &= E\{Y(t)\} = 1 \times P\{X(t) \le x\} + 0 \times P\{X(t) > x\} \\ &= F_X(x) \end{aligned}$$

15.12 CORRELATION FUNCTIONS

Autocorrelation Function and its Properties

We know that for WSS processes

1. $R_{xx}(\tau) = E[X(t)X(t+\tau)]$
   $\tau = t_2 - t_1$ with t_1, t_2 time assignments
   The following are properties of autocorrelation function:
   $$\left|R_{XX}(\tau)\right| \le R_{XX}(0)$$
   (i.e.,) $R_{XX}(\tau)$ is bounded by its value at the origin.
2. $R_{XX}(-\tau) = R_{XX}(\tau)$
   This shows that autocorrelation function has even symmetry.
3. $R_{XX}(0) = E[X^2(t)]$
   The bound at the origin is equal to mean squared value called the power in the process.
4. If $E[X(t)] = \tilde{X} \neq 0$ and $X(t)$ is ergodic with no periodic components then
   $$\underset{|\tau|\to\infty}{\text{Lt}} R_{XX}(\tau) = \tilde{X}^2$$
5. If $X(t)$ has a periodic component, then $R_{XX}(\tau)$ will have a periodic component with same period.
6. If $X(t)$ is ergodic, zero mean, and has no periodic component then
   $$\underset{|\tau|\to\infty}{\text{Lt}} R_{XX}(\tau) = 0$$
7. $R_{XX}(\tau)$ cannot have an arbitrary shape, that is, any arbitrary function cannot be an autocorrelation function.

Cross Correlation Function and its Properties

Some of the properties of cross correlation Function are as follows:

1. $R_{XY}(-\tau) = R_{YX}(\tau)$

The above property describes the symmetry of $R_{XY}(\tau)$.

2. $|R_{XY}(\tau)| \le \sqrt{R_{XX}(0)R_{YY}(0)}$, this gives a bound on the magnitude of $R_{XY}(\tau)$.

3. $|R_{XY}(\tau)| \le \frac{1}{2}[R_{XX}(0) + R_{YY}(0)]$.

Caution:

- Property 2 gives a tighter bound than property 3 since the geometric mean of two positive numbers cannot exceed their arithmetic mean.
- If $X(t)$ and $Y(t)$ are at least jointly WSS, $R_{XY}(t, t+\tau)$ is independent of absolute time and $R_{XY}(\tau) = E[X(t)Y(t+\tau)]$.
- If $R_{XY}(t, t+\tau) = 0$ then $X(t)$ and $Y(t)$ are called orthogonal processes.
- If the two processes are statistically independent the cross correlation function becomes $R_{XY}(t, t+\tau) = E[X(t)]E[Y(t+\tau)]$.
- In addition to being independent, $X(t)$ and $Y(t)$ are at least WSS, then $R_{XY}(\tau) = \tilde{X}\tilde{Y}$ (a constant).

15.13 COVARIANCE FUNCTIONS

The concept of covariance of two random variable defined in the earlier chapters can be extended to random processes.

The auto covariance function is defined as

$$C_{XX}(t, t+\tau) = E[\{X(t) - E[X(t)]\}\{X(t+\tau) - E[X(t+\tau)]\}] \tag{15.9}$$

This can also be expressed as

$$C_{XX}(t, t+\tau) = R_{XX}(t, t+\tau) - E[X(t)]E[X(t+\tau)]$$

The cross covariance function for two processes $X(t)$ and $Y(t)$ is defined as

$$C_{XY}(t, t+\tau) = E[\{X(t) - E[X(t)]\}\{Y(t+\tau) - E[Y(t+\tau)]\}] \tag{15.10}$$

This can also be expressed as

$$C_{XY}(t, t+\tau) = R_{XY}(t, t+\tau) - E[X(t)]E[Y(t+\tau)] \tag{15.10a}$$

For jointly WSS processes the above definitions are reduced to

$$C_{XX}(\tau) = R_{XX}(\tau) - \bar{X}^2 \tag{15.11}$$

and

$$C_{XY}(\tau) = R_{XY}(\tau) - \bar{X}\tilde{Y} \tag{15.12}$$

Caution:

- The variance of a random process is

$$\sigma_X^2 = E\{[X(t)] - E[X(t)]\}^2$$

This is obtained by putting $\tau = 0$ in equation (15.9).

- For a WSS process, variance does not depend on time which is given by $\tau = 0$ in equation (15.11).

$$\sigma_X^2 = E[\{X(t) - E[X(t)]\}]^2$$

$$= R_{XX}(0) - \tilde{X}^2$$

- Two processes are said to be uncorrelated

$$\text{if } C_{XY}(t, t+\tau) = 0$$

Hence, from equation (15.10a)

$$R_{XY}(t, t+\tau) = E[X(t)]E[Y(t+\tau)]$$

- Hence, two processes which are independent are uncorrelated. Its converse statement is not necessarily true.
- If the autocorrelation functions and the cross correlation functions of two random processes $X(t)$ and $Y(t)$ can be written in the form of a matrix as follows:

$$R(t_1, t_2) = \begin{bmatrix} R_{XX}(t_1, t_2) & R_{XY}(t_1, t_2) \\ R_{YX}(t_1, t_2) & R_{YY}(t_1, t_2) \end{bmatrix}$$

This matrix is called correlation matrix.

- If the correlation matrix can be written as

$$R(t_1 - t_2) = \begin{bmatrix} R_{XX}(t_1 - t_2) & R_{XY}(t_1 - t_2) \\ R_{YX}(t_1 - t_2) & R_{YY}(t_1 - t_2) \end{bmatrix}$$

Then the random processes $X(t)$ and $Y(t)$ are said to be jointly WSS.

15.14 SPECTRAL REPRESENTATION

We have so far studied the characteristics of random processes in the time domain. The characteristics of random process can be represented in the frequency domain. The autocorrelation function tells us something about how rapidly we can expect the random signal $X(t)$ to change as a function of time. If the autocorrelation function decays rapidly or slowly, it indicates that the process can be expected to change rapidly or slowly. Hence, the autocorrelation function contains information about the expected frequency content of the random process.

Power Density Spectrum

The spectral properties of a deterministic signal $x(t)$ are contained in its Fourier transform $X(\omega)$ which is given by

$$X(\omega) = \int_{-\infty}^{\infty} x(t) e^{-i\omega t} dt$$

The function $X(\omega)$ is called the spectrum of $x(t)$. Its units are volts per hertz.

If $X(\omega)$ is known, $x(t)$ can be recovered by means of Inverse Fourier transform

$$x(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} X(\omega) e^{i\omega t} d\omega$$

$X(\omega)$ may not exist for most sample functions of the process. Thus a spectral description of a random process utilizing a voltage density spectrum, that is, Fourier transform is not feasible because such a spectrum may not exist.

Definition

For a random process $\{X(t)\}$ let $x_T(t)$ be defined as a portion of the sample function $x(t)$ which exists between $-T$ and T

(i.e.,)
$$X_T(t) = \begin{cases} x(t), & -T < t < T \\ 0, & \text{otherwise} \end{cases}$$

The Fourier transform of this function will be $X_T(\omega) = \int_{-T}^{T} x_T(t) e^{-i\omega t} dt$

$$= \int_{-T}^{T} x(t) e^{-i\omega t} dt \tag{15.13}$$

The energy contained in $x(t)$ in the interval $(-T, T)$

$$E(T) = \int_{-T}^{T} x_T^2(\omega) dt = \int_{-T}^{T} x^2(t) dt$$

$$= \frac{1}{2\pi} \int_{-\infty}^{\infty} |X_T(\omega)|^2 \, d\omega$$

The average power $P(T)$ in $x(t)$ over $(-T, T)$ is

$$P(T) = \frac{1}{2T} \int_{-T}^{T} x^2(t) dt = \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{|X_T(\omega)|^2}{2T} d\omega$$

Thus average power P_{XX} in a random process $X(t)$ is given by the time average of its second moment:

$$P_{XX} = \underset{T \to \infty}{\text{Lt}} \frac{1}{2T} \int_{-T}^{T} E\left[X^2(t)\right] dt$$
$$= A\{E[X^2(t)]\}$$

For a process which is at least WSS,
$E[X^2(t)] = \bar{X}^2$, a constant and $P_{XX} = \bar{X}^2$.
Hence, average power P_{XX} contained in the random process is given by

$$P_{XX} = \frac{1}{2\pi} \int_{-\infty}^{\infty} \delta_{XX}(\omega) \, d\omega$$

where $\delta_{XX}(\omega)$ represents the power density spectrum or power spectral density of the process and is given by

$$\delta_{XX}(\omega) = \underset{T \to \infty}{\text{lt}} \frac{E\left[|X_T(\omega)|^2\right]}{2T} \tag{15.14}$$

Properties of Power Density Spectrum

The following are the properties of power density spectrum:

1. $\delta_{XX}(\omega) \geq 0$. Since expected value of a non-negative function is non-negative.

2. $\delta_{XX}(-\omega) = \delta_{XX}(\omega)$ for real $X(t)$.
3. $\delta_{XX}(\omega)$ is a real since $\left|X_{\Gamma}(\omega)\right|^2$ is real.
4. $\dfrac{1}{2\pi}\int_{-\infty}^{\infty} \delta_{XX}(\omega)d\omega = A\{E[X^2(t)]\}$
5. $\delta_{XX}(\omega) = \omega^2 \delta_{XX}(\omega)$

   (i.e.,) the power density spectrum of the derivatives $X(t) = \dfrac{d}{dt}X(t)$ is ω^2 times the power spectrum of $X(t)$.
6. $\dfrac{1}{2\pi}\int_{-\infty}^{\infty} \delta_{XX}(\omega)e^{i\omega t}d\omega = A[R_{XX}(t, t+\tau)]$

$$\delta_{XX}(\omega) = \int_{-\infty}^{\infty} A\, R_{XX}(t, t+\tau)e^{-i\omega t}dt$$

   (i.e.,) The power density of spectrum and the time average of the autocorrelation function forms a Fourier transform pair.
7. If $X(t)$ is at least WSS, $A[R_{XX}(t, t+\tau)] = R_{XX}(\tau)$. The above property becomes

$$\delta_{XX}(\omega) = \int_{-\infty}^{\infty} R_{XX}(\tau)e^{-i\omega t}dt$$

$$R_{XX}(\omega) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \delta_{XX}(\omega)e^{i\omega t}d\omega$$

Cross Power Density Spectrum

Let $X(t)$ and $Y(t)$ be two jointly WSS random processes with their cross correlation functions given by $R_{XY}(\tau)$ and $R_{YX}(\tau)$.

$$\delta_{XY}(\omega) = \int_{-\infty}^{\infty} R_{XY}(\tau)e^{-i\omega t}dt$$

where

$$R_{XY}(\tau) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \delta_{XY}(\omega)e^{iwt}d\omega$$

$$\delta_{YX}(\omega) = \int_{-\infty}^{\infty} R_{YX}(t)e^{-i\omega t}dt$$

where

$$R_{YX}(\tau) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \delta_{YX}(\omega)e^{i\omega t}d\omega$$

For two real random processes $X(t)$ and $Y(t)$, we define $X_T(t)$ and $Y_T(t)$ as truncated ensemble members.

(i.e.,)

$$X_T(t) = \begin{cases} x(t), & -T < t < T \\ 0, & \text{elsewhere} \end{cases}$$

and

$$Y_T(t) = \begin{cases} y(t), & -T < t < T \\ 0, & \text{elsewhere} \end{cases}$$

Let $X_T(\omega)$ and $Y_T(\omega)$ be Fourier transforms of $x_T(t)$ and $y_T(t)$, respectively.
The cross power $P_{XY}(T)$ in the two processes within the interval $(-T, T)$ by

$$P_{XY}(T) = \frac{1}{2T}\int_{-T}^{T} x_T(t)\,y_T(t)\,dt$$

$$= \frac{1}{2T}\int_{-T}^{T} x(t)\,y(t)\,dt$$

By using Parseval's theorem

$$P_{XY}(T) = \frac{1}{2T}\int_{-T}^{T} x(t)\,y(t)\,dt = \frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{X_T^*(\omega)Y_T(\omega)\,d\omega}{2T}$$

The average cross power P_{XY} in the random processes $X(t)$ and $Y(t)$ is given by

$$P_{XY} = \underset{T\to\infty}{\text{Lt}}\left\{\frac{1}{2\pi}\int_{-\infty}^{\infty}\frac{E\left[X_T^*(\omega)Y_T(\omega)\right]}{2T}\right\}d\omega$$

$$= \frac{1}{2\pi}\int_{-\infty}^{\infty}\left\{\underset{T\to\infty}{Lt}\,\frac{E\left[X_T^*(\omega)Y_T(\omega)\right]}{2T}\right\}d\omega$$

$$= \frac{1}{2\pi}\int_{-\infty}^{\infty}\delta_{XY}(\omega)d\omega$$

where $\delta_{XY}(\omega)$ represents the cross density spectrum or cross spectral density of the processes $X(t)$ and $Y(t)$ and is given by

$$\delta_{XY}(\omega) = \underset{T\to\infty}{\text{Lt}}\,\frac{E[X_T^*(\omega)Y_T(\omega)]}{2T}$$

Similarly the average cross power is given by

$$P_{YX} = \frac{1}{2\pi}\int_{-\infty}^{\infty}\delta_{YX}(\omega)d\omega$$

where $\delta_{YX}(\omega) = \underset{T\to\infty}{\text{Lt}}\,\dfrac{E\left[Y_T^*(\omega)X_T(\omega)\right]}{2T}$

Properties of Cross Power Density Spectrum

Some of the properties of cross power Spectrum of real random processes $X(t)$ and $Y(t)$ are as follows:

1. $\delta_{XY}(\omega) = \delta_{YX}(-\omega) = \delta_{YX}^*(\omega)$.
2. The real parts of $\delta_{XY}(\omega)$ and $\delta_{YX}(\omega)$ are even functions of ω.
3. The Imaginary parts of $\delta_{XY}(\omega)$ and $\delta_{YX}(\omega)$ are odd functions of ω.
4. If $X(t)$ and $Y(t)$ are orthogonal then $\delta_{XY}(\omega) = 0$ and $\delta_{YX}(\omega) = 0$.
5. If $X(t)$ and $Y(t)$ are uncorrelated and have constant means $\bar{X}$ and $\bar{Y}$, then

$$\delta_{XY}(\omega) = \delta_{YX}(\omega) = 2\pi\bar{X}\,\bar{Y}\,\delta(\omega).$$

Relation Between Cross Power Spectrum and Cross Correlation Function

Consider $X_T(\omega) = \int_{-T}^{T} X(t)e^{-i\omega t}dt$

$$Y_T(\omega) = \int_{-T}^{T} Y(t_1)e^{-i\omega t_1}dt_1$$

We can form

$$X_T^*(\omega)Y_T(\omega) = \int_{-T}^{T} X(t)e^{i\omega t}dt \int_{-T}^{T} Y(t_1)e^{-i\omega t_1}dt_1$$

The cross power density spectrum is given by

$$\delta_{XY}(\omega) = \underset{T\to\infty}{\text{Lt}} \frac{E\left[X_T^*(\omega)Y_T(\omega)\right]}{2T}$$

$$= \underset{T\to\infty}{\text{Lt}} \frac{1}{2T} E\left[\int_{-T}^{T} X(t)e^{i\omega t}dt \int_{-T}^{T} Y(t_1)e^{-i\omega t_1}dt_1\right]$$

$$= \underset{T\to\infty}{\text{Lt}} \frac{1}{2T} \int_{-T}^{T}\int_{-T}^{T} R_{XY}(t,t_1)e^{-i\omega(t_1-t)}dt\,dt_1$$

$$\therefore \frac{1}{2\pi}\int_{-\infty}^{\infty} \delta_{XY}(\omega)e^{i\omega t}d\omega = \frac{1}{2\pi}\int_{-\infty}^{\infty} \underset{T\to\infty}{\text{Lt}} \frac{1}{2T} \int_{-T}^{T}\int_{-T}^{T} R_{XY}(t,t_1)e^{-i\omega(t_1-t)}dt\,dt_1 e^{i\omega t}d\omega$$

$$= \underset{x\to\infty}{\text{Lt}} \frac{1}{2T} \int_{-T}^{T}\int_{-T}^{T} R_{XY}(t,t_1)\frac{1}{2\pi}\int_{-\infty}^{\infty} e^{i\omega(t-t_1+t)}d\omega\; dt_1\,dt$$

$$= \underset{x\to\infty}{\text{Lt}} \frac{1}{2T} \int_{-T}^{T}\int_{-T}^{T} R_{XY}(t,t_1)\delta(t_1-\tau-t)\,dt_1\,dt$$

$$\therefore \frac{1}{2\pi}\int_{-\infty}^{\infty} \delta_{XY}(\omega)e^{i\omega t}d\omega = \underset{T\to\infty}{\text{Lt}} \frac{1}{2T} \int_{-T}^{T} R_{XY}(t,t+\tau)\,dt$$

This is valid $-T < t + \tau < T$

Relation Between Power Spectrum and Autocorrelation Function

The Fourier transform of $x(t)$ is given by

$$X_T(\omega) = \int_{-T}^{T} x(t)e^{-i\omega t}dt \tag{15.15}$$

The Power density spectrum for the random process is given by

$$\delta_{XX}(\omega) = \underset{T\to\infty}{\text{Lt}} \frac{E[|X_T(\omega)|^2]}{2T} \tag{15.16}$$

Since $|X_T(\omega)|^2$ is real

$$|X_T(\omega)|^2 = X_T(\omega)X_T(-\omega)$$

$$= \int_{-T}^{T} X(t_1)e^{i\omega t_1}dt_1 \int_{-T}^{T} X(t_2)e^{-i\omega t_2}dt_2$$

Substituting equations (15.15) and (15.16),

$$\delta_{XX}(\omega) = \underset{T\to\infty}{\text{Lt}}\, E\left[\frac{1}{2T}\int_{-T}^{T} X(t_1)e^{i\omega t_1}dt_1 \int_{-T}^{T} X(t_2)e^{-i\omega t_2}dt_2\right]$$

$$= \underset{T\to\infty}{\text{Lt}}\frac{1}{2T}\int_{-T}^{T}\int_{-T}^{T} E[X(t_1)X(t_2)]e^{-i\omega(t_1-t_2)}dt_2\, dt_1$$

The expectation in the integral is the autocorrelation function of $X(t)$.

$$E[X(t_1)X(t_2)] = R_{XX}(t_1, t_2)$$

where $-T < (t_1, t_2) < T$

$$\therefore \delta_{XX}(\omega) = \underset{T\to\infty}{\text{Lt}}\frac{1}{2T}\int_{-T}^{T}\int_{-T}^{T} R_{XX}(t_1, t_2)e^{-i\omega(t_1-t_2)}dt_1dt_2$$

The inverse transform is

$$\frac{1}{2\pi}\int_{-\infty}^{\infty}\delta_{XX}(\omega)e^{i\omega\tau}d\omega = \frac{1}{2\pi}\int_{-\infty}^{\infty}\delta_{XX}(\omega)e^{i\omega\tau}d\omega$$

$$= \frac{1}{2\pi}\int_{-\infty}^{\infty}\underset{T\to\infty}{\text{Lt}}\frac{1}{2T}\int_{-T}^{T}\int_{-T}^{T} R_{XX}(t,t_1)e^{-i\omega(t_1,-t)}dt\, dt_1 e^{i\omega\tau}d\omega$$

$$= \underset{T\to\infty}{\text{Lt}}\frac{1}{2\pi}\int_{-T}^{T}\int_{-T}^{T} R_{XX}(t,t_1)\frac{1}{2\pi}\int_{-\infty}^{\infty} e^{i\omega(\tau-t_1+t)}d\omega\, dt_1 dt$$

$$\underset{T\to\infty}{\text{Lt}}\int_{-T}^{T} e^{i\omega(\tau+t-\tau+t)}dt$$

If the random process $X(t)$ is at least WSS $< \mu_{XX} R_{XX}(t, t+\tau) >= R_{XX}(t)$

$$\delta_{XX}(\omega) = \int_{-\infty}^{\infty} R_{XX}(t)e^{-i\omega t}dt$$

$$R_{XX}(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty}\delta_{XX}(\omega)e^{i\omega t}d\omega$$

The above relations are called Wiener-Khintchine relations after the great American Mathematician Norbert Wiener and Russian Mathematician A. I. Khinchine.

They give the relation between time domain description, that is, correlation function of processes and their description in the frequency domain, that is, power spectrum.

Worked Out Examples

EXAMPLE 15.2

Prove that the Random process $X(t) = A\cos(\omega_C t + \theta)$ is not stationary if it is assumed that ω_C and A are constants and θ is a uniformly distributed variable on the interval $(0, \pi)$.

Solution: According to the definition a random process is not stationary if mean value or the autocorrelation function of the random process is a function of time t.

$$E[X(t)] = \int_0^{\pi} \frac{1}{\pi} A\cos(\omega_C t + \theta)d\theta$$

$$= \frac{A}{\pi}\int_0^{\pi} \cos(\omega_C t + \theta)d\theta$$

$$= \frac{A}{\pi}\int_0^{\pi} (\cos\omega_C t\cos\theta - \sin\omega_C t\sin\theta)d\theta$$

$$= \frac{A}{\pi}\int_0^{\pi} \cos\omega_C t\cos\theta\, d\theta - \frac{A}{\pi}\int_0^{\pi} \sin\omega_C t\sin\theta\, d\theta$$

$$= \frac{A}{\pi}\cos\omega_C t\int_0^{\pi} \cos\theta\, d\theta - \frac{A}{\pi}\sin\omega_C t\int_0^{\pi} \sin\theta\, d\theta$$

$$= \frac{A}{\pi}\left[\cos\omega_C t(\sin\theta)\Big|_0^{\pi} - \sin\omega_C t(-\cos\theta)\Big|_0^{\pi}\right]$$

$$= \frac{A}{\pi}(-2\sin\omega_C t)$$

We can observe that this is a function of time t, hence $X(t)$ is not a stationary process.

Caution:

Hence $X(t)$ is WSS since autocorrelation function depends only on t and mean value is a constant, if $X(t) = A\cos(\omega_C t + \theta)$ where θ is uniformly distributed on the interval $(0, 2\pi)$.

EXAMPLE 15.3

Prove that $X(t)$ is an ergodic random process if $X(t) = \cos(\omega t + \theta)$ where ω is a constant and θ is a random variable with probability density $P(\theta) = \begin{cases} \frac{1}{2\pi}, & 0 \le \theta \le 2\pi \\ 0, & \text{elsewhere} \end{cases}$.

Solution: A random process is ergodic if the time averages and ensemble averages are equal. In the previous problem we have $A = 1$ and $\omega_C = \omega$

Hence the mean

$$E[X(t)] = \int_0^{2\pi} \frac{1}{2\pi}\cos(\omega t + \theta)\, d\theta$$

$$= \frac{1}{2\pi}\left|\sin(\omega t + \theta)\right|_0^{2\pi}$$

$$= \frac{1}{2\pi}[\sin \omega t - \sin \omega t] = 0 = \text{constant}$$

Auto correlation function,

$$R_{XX}(t, t+\tau) = E[X_t X_{t+\tau}]$$

$$= E\ (\cos(\omega t + \theta)\cos(\omega t + \omega\tau + \theta)$$

$$= E\left[\frac{\cos(\omega t + \omega t + \omega\tau + \theta + \theta) + \cos(\omega t + \theta - \omega t - \omega\tau - \theta)}{2}\right]$$

$$\because \cos A \cos B = \frac{\cos(A+B) + \cos(A-B)}{2}$$

$$= \frac{1}{2} E[\cos(2\omega t + \omega\tau + 2\theta) + \cos(-\omega\tau)]$$

$$= \frac{1}{2}\left[\int_0^{2\pi} \frac{1}{2\pi}\cos(2\omega t + \omega\tau + 2\theta)d\theta + \int_0^{2\pi} \frac{1}{2\pi}\cos(\omega\tau)d\theta\right]$$

$$= \frac{1}{4\pi}\left[\frac{\sin 2\omega\tau + \omega\tau + 2\theta}{2}\Big|_0^{2\pi} + \frac{1}{2}\cos\omega t = \frac{1}{2}\cos\omega t\right.$$

The time average can be determined by

$$< X(t) > = \underset{T\to\infty}{\text{Lt}} \frac{1}{2T}\int_{-T}^{T} \cos(\omega t + \theta)dt$$

$$= \underset{T\to\infty}{\text{Lt}} \frac{1}{2T}\left|\frac{\sin\omega t + \theta}{\omega}\right|_{-T}^{T} = \underset{T\to\infty}{\text{Lt}} \frac{1}{2\text{T}}[\sin(\omega T + \theta) - \sin(-\omega T + \theta)]$$

As Lt $T \to \infty$, $< X(t) > = 0$.
The time autocorrelation function of the process can be determined by

$$R_{XX}(\tau) = < X(t)\ X(t+\ \tau) >$$

$$= \underset{T\to\infty}{\text{Lt}} \frac{1}{2T}\int_{-T}^{T} \cos(\omega t + \theta)\cos(\omega t + \omega\tau + \theta)dt$$

$$= \underset{T\to\infty}{\text{Lt}} \frac{1}{2T}\int_{-T}^{T} \frac{\cos(2\omega t + \omega\tau + 2\theta) + \cos\omega\tau}{2}dt$$

$$= \underset{T\to\infty}{\text{Lt}} \frac{1}{4T}\left[\left|\frac{\sin(2\omega t + \omega\tau + 2\theta)}{2\omega}\right|_{-T}^{T} + \left|t\cos\omega\tau\right|_{-T}^{T}\right]$$

$$= \underset{T\to\infty}{\text{Lt}} \frac{1}{4T}[0 + 2T\cos\omega\tau]$$

$$= \frac{\cos\omega\tau}{2}$$

Hence, we can observe that the ensemble averages and time averages are equal.
Hence the given random process is an ergodic random process.

EXAMPLE 15.4

Find the mean and variance of the process $X(t)$ given that the autocorrelation function for a stationary ergodic process with no periodic components is $R_{XX}(\tau) = 25 + \frac{4}{1+6\tau^2}$.

Solution: By the property of autocorrelation function,

$$\mu_x^2 = \underset{\tau \to \infty}{\text{Lt}}\, R_{XX}(\tau)$$

$$= 25$$

$$\therefore \mu_x = 5$$

$$E\{X^2(t)\} = R_{XX}(0)$$

$$= 25 + 4$$

$$= 29$$

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

$$= 29 - 25 = 4$$

EXAMPLE 15.5

Consider a random process $X(t)$ whose mean is zero and $R_{XX}(\tau) = A\delta(\tau)$ show that the process $X(t)$ is mean-ergodic.

Solution: Given the mean $E[X(t)] = 0$.
Its covariance is equal to the autocorrelation function

$$C_{XX}(\tau) = R_{XX}(\tau) = A\delta(\tau)$$

$$\sigma_T^2 = \frac{1}{T}\int_0^T A\delta(\tau)\left[1 - \frac{|\tau|}{2T}\right]d\tau$$

$$= \frac{A}{T}\int_0^T \delta(t)\left[1 - \frac{|\tau|}{2T}\right]d\tau$$

$$= \frac{A}{T}$$

$$\sigma_T = \sqrt{\frac{A}{T}} \quad \text{where } T \to \infty, \sigma_T \to 0$$

$\therefore X(t)$ is mean-ergodic.

EXAMPLE 15.6

Given that the random process $X(t) = 10\cos(100t + \phi)$ where ϕ is a uniformly distributed random variable in the interval $(-\pi, \pi)$. Show that the process is correlation ergodic.

Solution: Given that $X(t) = 10\cos(100t + \phi)$
We have to prove that when $T \to \infty$

$$\frac{1}{2T}\int_{-T}^{T} X(t)X(t+\tau)dt \to R_{XX}(\tau)$$

$$R_{XX}(\tau) = E[X(t)X(t+\tau)]$$

$$R_{XX}(\tau) = E\{10\cos(100t+\phi)10\cos\{100(t+\tau)+\phi\}]$$

$$= E\{100\left[\frac{\cos(200t+100\tau+2\phi)+\cos(100\tau)}{2}\right]$$

$$= 50\,E[\cos(200t+100\tau+2\phi)]+50\,E[\cos 100\tau]$$

$$= 0+50E[\cos 100\tau] = 50\cos 100\tau \qquad (15.17)$$

Now,

$$\frac{1}{2T}\int_{-T}^{T} 10\cos(100t+\phi)10\cos[100(t+\tau)+\phi]dt$$

$$= \frac{1}{2\mathrm{T}}\int_{-\mathrm{T}}^{\mathrm{T}} 50\cos 100\tau\,\mathrm{dt} = \frac{50}{2\mathrm{T}}\cos 100\tau\,|\mathrm{t}|_{-T}^{T}$$

$$= 50\cos 100\tau \qquad (15.18)$$

Since (15.17) and (15.18) are equal, $X(t)$ is correlation ergodic.

EXAMPLE 15.7

The random binary transmission process $X(t)$ is a WSS process with zero mean and autocorrelation function $R(\tau) = 1 - \frac{|\tau|}{T}$, where T is a constant. Find the mean and variance of the time average of $X(t)$ over (0, T). Is $X(t)$ mean-ergodic?

Solution:

$$\bar{X}_T = \frac{1}{T}\int_0^T X(t)dt$$

$$E(\bar{X}_T) = E(X(t)) = 0$$

$$\mathrm{Var}(\bar{X}_T) = \frac{1}{T}\int_0^T \left\{1-\frac{|\tau|}{T}\right\}^2 d\tau$$

$$= \frac{2}{T}\int_0^T \left\{1-\frac{\tau}{T}\right\}^2 d\tau = \frac{2}{T}\left|\frac{(1-\frac{\tau}{T})^3}{-3T}\right|_0^T$$

$$= \frac{2}{3}$$

$$\underset{T\to\infty}{\text{Lt}}\,[\text{var}(\bar{X}(t))] = \frac{2}{3} \neq 0$$

(i.e.,) Condition for mean ergodicity of $X(t)$ is not satisfied. Hence, $X(t)$ is not mean-ergodic.

Work Book Exercises

1. Define autocorrelation function of a stationary process.
2. Find the mean of the stationary process $X(t)$, whose autocorrelation function is given by

$$R_{XX}(\tau) = \frac{25\tau^2 + 36}{6.25\tau^2 + 4}.$$

3. Find the variance of the stationary process $X(t)$ whose autocorrelation function is given by

$$R_{XX}(\tau) = 16 + \frac{9}{1 + 6\tau^2}.$$

4. Define ensemble average and time average of a random process (stochastic process) $X(t)$.

15.15 DISCRETE TIME PROCESSES

Let us first define a band-limited process before going to discrete time processes. A band-limited process is one with a power density spectrum that is zero at all frequencies except over a finite band where $|\omega| < \omega_s/2$ with $\omega_s = \frac{2\pi}{T_s}$ and T_s is the constant time between any adjacent pairs of samples. T_s is called the sampling interval or sampling period.

Let $X(t)$ be a band limited random process for which the samples are taken at times nT_s, $n = 0, \pm1, \pm2, \ldots$ to form a discrete time process $X(nT_s)$.

Here, sampling is the result of multiplying $X(t)$ by a periodic sampling pulse train consisting of rectangular pulses of short duration T_p and amplitude $\frac{1}{T_p}$ each occurring for T_s seconds. Then the autocorrelation function $R_{XX}(\tau)$ has a sampled representation, denoted by $R_{X_sX_s}(\tau)$ and is given by,

$$R_{X_sX_s}(\tau) = R_{XX}(\tau) \sum_{n=-\infty}^{\infty} \delta(\tau - nT_s)$$

where δ is an impulse function defined as follows:

If $\phi(x)$ is any arbitrary function of x, $x_1 < x_2$ are any two values of x, x_0 is the point of occurrence of impulse then $\delta(x)$ is

$$\int_{x_1}^{x_2} \phi(x)\,\delta(x - x_0)\,dx = \begin{cases} 0, & x_2 < x_0 \quad \text{or} \quad x_0 < x_1 \\ \frac{1}{2}[\phi(x_0^+) + \phi(x_0^-)], & x_1 < x_0 < x_2 \\ \frac{1}{2}\phi(x_0^+), & x_0 = x_1 \\ \frac{1}{2}\phi(x_0^-), & x_0 = x_2 \end{cases}$$

$$\therefore R_{X_S X_S}(\tau) = \sum_{n=-\infty}^{\infty} R_{XX}(nT_s)\delta(\tau - nT_s) \tag{15.19}$$

Direct Fourier transformation of the above equation (15.19) gives the power spectrum of the discrete time random process given by,

$$\delta_{X_S X_S}(\omega) = \sum_{n=-\infty}^{\infty} R_{XX}(nT_s)e^{-in\omega T_s} \tag{15.20}$$

15.16 DISCRETE TIME SEQUENCES

Here simple notations are changed since computers treat samples as simply a sequence of numbers but do not mind about the separation of time T_s which is used between samples.

Hence, the explicit dependence on T_s is dropped and $X(nT_s)$ is written as $X[n]$, a function of n.

Let Ω denote the discrete frequency given $\Omega = \omega T_s$

Here Z-transforms are extensively used.

The R.H.S of equation (15.21) is a Z-transform of the sequence of samples of the autocorrelation function.

$\therefore$ The two-sided Z-transform denoted by $\delta_{X_s X_s}(Z)$ of the sequence $R_{XX}[n]$ is

$$\delta_{X_S X_S}(Z) = \sum_{n=-\infty}^{\infty} R_{XX}[n]Z^{-n} \tag{15.21}$$

By comparing equations (15.18) and (15.19) we have

$$\begin{aligned}\delta_{X_S X_S}(\omega) &= \sum_{n=-\infty}^{\infty} R_{XX}[n]e^{-in\omega T_s} \\ &= \sum_{n=-\infty}^{\infty} R_{XX}[n]e^{-in\Omega} \\ &= \delta_{X_S X_S}(Z)\Big|_{Z=e^{i\Omega}} \\ &= \delta_{X_S X_S}(e^{i\Omega})\end{aligned}$$

15.17 SOME NOISE DEFINITIONS

White Noise

A sample function $n(t)$ of a WSS noise random process $N(t)$ is called white noise if the power density spectrum of $N(t)$ is a constant at all frequencies.

$$\delta_{NN}(\omega) = \frac{No}{2_{(u)}} \text{ for white noise.}$$

where N_o is a real positive constant. The autocorrelation function of $N(t)$ is the inverse Fourier transformation of equation given by

$$R_{NN}(\tau) = \left(\frac{No}{2}\right)\delta(\tau)$$

White noise is unrealizable since it possesses infinite average power.

$$\frac{1}{2\pi}\int_{-\infty}^{\infty} \delta_{NN}(\omega)d\omega = \infty$$

Coloured Noise

Any noise that is not white is coloured. If light has only a portion of the visible light frequencies in its spectrum.

Caution:

- Thermal noise which is generated by thermal agitation of electrons in any electrical conductor has a power spectrum that is constant up to very high frequencies and then decreases.
- Noise having a non-zero and constant power spectrum over a finite frequency band and zero everywhere else is called band-limited white noise.

15.18 TYPES OF NOISE

The definitions of noise given in the previous section characterize noise through its power density spectrum. While finding the response of a linear system, when a random waveform is applied at its input, the system is assumed to be free of any internally generated noise. We have to model techniques for both the network and for the external source that drives the network, all the internally generated network noise can be thought of as having been caused by the external source. Some noise sources are as follows:

Resistive (Thermal) Noise Source

Consider an ideal voltmeter which is noise free and finite impedance that responds to voltages with a small ideal frequency band $\frac{d\omega}{2\pi}$ and at an angular frequency ω.

A noise voltage $e_n(t)$ exists when this voltmeter is used to measure the voltage across a resistor of resistance R ohms and which has mean squared value as

$$\overline{e_n^2(t)} = \frac{2KTR\,d\omega}{\pi} \tag{15.22}$$

Here, $K = 1.38 \times 10^{-23}$ joules per kelvin is Boltzmann's constant and T is temperature in kelvin. As the voltmeter does not load the resistor, $\overline{e_n^2(t)}$ is the mean squared open circuit voltage of the resistor which can be treated as a voltage source with internal impedance R.

Hence, the noisy resistor can be modelled as Thevenin voltage source, whose short-circuit mean squared current is

$$\overline{i_n^2(t)} = \frac{\overline{e_n^2(t)}}{R^2} = \frac{2KT\,d\omega}{\pi R} \tag{15.23}$$

In addition, the incremental noise power dN_L, delivered to the load in the incremental band $d\omega$ by the noisy resistor as a source is

$$dN_L = \frac{\overline{e_n^2(t)}R_L}{(R+R_L)^2} = \frac{2KTRR_L d\omega}{\pi(R+R_L)^2} \tag{15.24}$$

The incremental available power of the source is the maximum power which occurs when $R_L = R$ and is given by

$$dN_{as} = \frac{\overline{e_n^2(t)}}{4R} = \frac{KTd\omega}{2\pi} \tag{15.25}$$

Caution:

- This incremental power available from a resistor source is independent of the resistance of the source.
- It depends only on its physical temperature *T*.

Shot Noise

The term "shot noise" can be used to describe any noise source. For instance, particle simulation may produce a certain amount of noise where due to the small number of particles simulated, the simulation exhibits undue statistical fluctuations which don't reflect the real-world system.

The magnitude of shot noise increases according to the square root of the expected number of events, such as electrical current or intensity of light.

Since the strength of the signal increases more rapidly, the relative proportion of shot noise decreases and Signal to Noise Ratio (SNR) increases.

The SNR is given by,

$$\text{SNR} = \sqrt{N}\text{, where } N \text{ is the average number of events.}$$

Given a set of Poisson points t_i, with average density λ, a real function $h(t)$, we form the sum

$$s(t) = \sum_i h(t - t_i)$$

The mean μ_s and variance σ_s^2 of the shot noise process $s(t)$ are

$$\mu_s = \lambda \int_{-\infty}^{\infty} h(t)\,dt,\ \sigma_s^2 = \lambda \int_{-\infty}^{\infty} h^2(t)dt$$

Caution:

- Poisson points are realistic models for a large class of electron emissions, telephone calls, date communications, visits to a doctor, arrivals at a park, etc.
- When *N* is very large SNR is very large and any relative fluctuations in *N* due to other sources are more likely to dominate over shot noise.
- When thermal noise is at fixed level, an increase in *N* leads to dominance of shot noise.

Extraterrestrial Noise

Noises obtained from sources other than those related to the earth are as follows:

(a) **Solar noise:** This is the noise that originates from the sun. The sun radiates a broad spectrum of frequencies, including those which are used for broadcasting. The sun is an active star which is constantly changing. It undergoes cycles of peak activity from which electrical disturbances erupt. This cycle is about 11 years long.

(b) **Cosmic noise:** Distant stars also radiate noise as much as the way the sun does. The noise received from them is called block body noise. Noise also cones from distant galaxies in the same way as they come from the Milky Way.

Extraterrestrial noise is observable at frequencies in the range from about 8 MHz to 1.43 GHz. Apart from man-made noise, it is strongest component over the range of 20 to 120 MHz, and not much of it below 20 MHz penetrates below the ionosphere.

Arbitrary Noise Sources and Effective Noise Temperature

Let us suppose that an actual noise source has an incremental available noise power dN_{as} open circuit output mean square voltage $\overline{e_n^2(t)}$ and the impedance measured between its output terminals of $z_0(\omega) = R_0(\omega) + j\,X_0(\omega)$.

Then the available noise power is

$$dN_{as} = \frac{\overline{e_n^2(t)}}{4R_o(\omega)} \tag{15.26}$$

From equation (15.22), if we define an effective noise temperature T_s by attributing all the sources noise to the resistive part $R_o(\omega)$ of its output impedance, then

$$\overline{e_n^2(t)} = 2KT_sR_o(\omega)\frac{d\omega}{\pi} \tag{15.27}$$

The available power is still independent of the source impedance but depends on the source's temperature.

$$dN_{as} = KT_s\frac{d\omega}{2\pi} \tag{15.28}$$

with a purely resistance source.

Average Noise Figures

The frequency band is not incremental in a realistic network. Hence, some quantities such as noise temperature and noise figure are not constant but they become frequency dependent.

Average operating noise figure denoted by F_{op} is defined as the total output available noise power Na_o from a network divided by the total output available noise power N_{aos} due to the source.

$$\therefore \bar{F}_{op} = \frac{N_{ao}}{N_{aos}}$$

If G_a is the network's available power gain, the available output noise power due to the source alone is

$$dN_{aos} = G_a dN_{as} = G_a KT_s\frac{d\omega}{2\pi}$$

$$\therefore N_{aos} = \frac{K}{2\pi}\int_0^\infty T_s G_a d\omega$$

Similarly, effective input noise temperature T_e of a network is a measure of its noise performance. Another measure of performance is incremental or spot noise figure denoted by F given by,

$$F = \frac{dN_{ao}}{dN_{aos}} = \frac{\text{Total incremental available noise power}}{\text{Incremental available output noise power}}$$

$$= 1 + \frac{T_e}{T_s}$$

Whenever we define standard source having standard noise temperature $T_o = 290\ K$, standard spot noise figure, F_o is given by

$$F_o = 1 + \frac{T_e}{T_o}$$

When a network is used with the source for which it is intended to operate F will be operating spot noise figure F_{op} given by,

$$F_{op} = 1 + \frac{T_e}{T_s}.$$

$$\text{Hence, } \bar{F}_{op} = \frac{\int_0^\infty F_{op} T_S G_a d\omega}{\int_0^\infty T_s G_a d\omega}$$

Caution:

- If the source's temperature is approximately constant, operating average noise figure becomes,

$$\bar{F}_{op} = \frac{\int_0^\infty F_{op} G_a d\omega}{\int_0^\infty G_a d\omega}$$

- If the source is standard then average standard noise figure $\bar{F}_o$ becomes

$$\bar{F}_o = \frac{\int_0^\infty F_o G_a d\omega}{\int_0^\infty G_a d\omega}$$

Average Noise Temperature

The incremental available output noise power from a network with available power gain G_a that is driven by a source of temperature T_s and effect input noise temperature T_e is

$$dN_{ao} = G_a K (T_s + T_e) \frac{d\omega}{2\pi}$$

$\therefore$ Total available power $N_{ao} = \int_0^\infty dN_{ao}$

$$= \frac{K}{2\pi} \int_0^\infty G_a (T_s + T_e) d\omega$$

In addition, average effective source temperature $\bar{T}_s$ and average effective input noise temperature $\bar{T}_e$ are constant temperatures that produce the same total available power.

$$N_{ao} = \frac{K}{2\pi} \left(\bar{T}_s + \bar{T}_e \right) \int_0^\infty G_a d\omega$$

$$\therefore \bar{T}_s = \frac{\int_0^\infty T_s G_a d\omega}{\int_0^\infty G_a d\omega}$$

$$\text{and } \bar{T}_e = \frac{\int_0^\infty T_e G_a d\omega}{\int_0^\infty G_a d\omega}$$

In addition, noise band width of a network is given by,

$$\omega_N = \frac{\int_0^\infty G_a(\omega)d\omega}{G_a(\omega_0)}$$

where ω_0 is the central band angular frequency of the function $G_a(\omega)$ and it is the available power gain.

Worked Out Examples

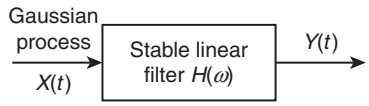
EXAMPLE 15.8

A Gaussian random process $X(t)$ is applied to a stable linear filter, show that the random process $Y(t)$ developed at the output of the filter is also Gaussian.

Solution: A Gaussian random process $X(t)$ is applied to a stable linear filter $H(\omega)$ as shown,

Stable filter must produce a finite mean square value, for the output process $Y(t)$ for all t. The pdf of the input process can be written as

$$f(x) = \frac{1}{\sqrt{2\pi\sigma_x^2}} e^{-\left[\frac{(x-\mu_x)^2}{2\sigma^2}\right]}$$



Stable linear filter

The output process of linear filter is finite.

$$y(t) = \int_0^T h(t-\tau)X(\tau)d\tau \qquad 0 \le t \le \infty$$

To demonstrate that $y(t)$ is also Gaussian, it is sufficient to show that any linear function defined on $Y(t)$ is also a Gaussian. We define this random variable.

$$Z = \int_0^\infty g_y(t)Y(t)dt$$

Let the mean square value of Z be

$$\begin{aligned} Z &= \int_0^\infty g_y(t) \cdot \left\{ \int_0^T h(t-\tau)X(\tau)d\tau \right\} d\tau \qquad 0 \le t \le \infty \\ &= \int_0^T X(\tau) \left\{ \int_0^\infty g_y(t)h(t-\tau)dt \right\} d\tau \\ &= \int_0^T X(\tau)g(\tau)d\tau \end{aligned}$$

Since $X(t)$ is a Gaussian process, Z is a Gaussian random variable. Thus, a stable linear filter always produces Gaussian output whenever its input is a Gaussian process.

EXAMPLE 15.9

A single $f(t) = e^{-4t}\,u(t)$ is passed through an ideal low pass filter with cutoff frequency 1 radian per second. Determine the energies of the input signal and the output signal.

Solution:

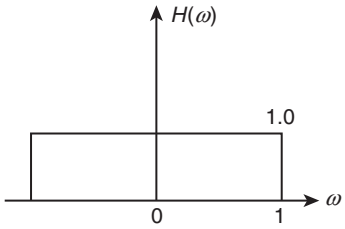
$$f(t) = e^{-4t}u(t)$$

$$F(\omega) = \int_{-\infty}^{\infty} e^{-4t}u(t)e^{-j\omega t}dt$$

$$= \int_{0}^{\infty} e^{-(4-j)t}dt = \frac{e^{-(4tj\omega)t}}{-(4+j\omega)}\Bigg|(t \text{ changes } 0 \text{ to to} \infty)$$

$$F(\omega) = \frac{1}{4+j\omega}$$

$$H(\omega) = 1 \quad -1 \le \omega \le 1$$



Spectrum of output signal $= F(\omega) \cdot H(\omega)$

$$= \frac{1}{4+j\omega} \quad \text{for} -1 \le \omega \le 1$$

$$= 0, \quad \text{otherwise}$$

Energy of input signal, $E_i = \int_{-\infty}^{\infty} f^2(t)dt$

$$= \int_{-\infty}^{\infty} \left\{e^{-4t}u(t)^2\right\}dt = \int_{0}^{\infty} e^{-8t}dt$$

$$= \frac{e^{-8t}}{-8}\Bigg|(t \text{ changes } 0 \text{ to} \infty) = \frac{1}{8} = 0.125$$

Energy of output signal $E_0 = \int_{-\infty}^{\infty} |F_0(\omega)|^2 d\omega$

$$= \int_{-1}^{1} \left|\frac{1}{4+j\omega}\right|^2 d\omega$$

$$= \int_{-1}^{1} \frac{1}{16+\omega^2} d\omega$$

$$= 2\int_{0}^{1} \frac{1}{16+\omega^2} d\omega$$

$$= 2.2\pi \frac{1}{2\pi \times 4} \cdot \tan^{-1}\frac{1}{4}$$

$$= \frac{1}{2}\tan^{-1}\frac{1}{4}$$

$$= 0.1225$$

EXAMPLE 15.10

Write short notes on the following:

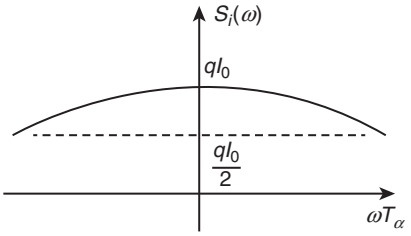
(i) Noise power spectral density

(ii) Noise suppression in semiconductor devices.

Solution:

(i) **Noise power spectral density:** Power spectral density (PSD) is defined as the power per unit bandwidth, measured in watts/Hz. White noise is the noise signal, whose PSD is uniform over the entire frequency range.

$$G_n(f) = \frac{n}{2}, \quad \text{for} -\infty < f < \infty$$



Noise power spectral density

Shot noise is present in both vacuum tubes and semiconductor devices. In vacuum tubes, shot noise arises due to random emission of electrons from the cathode. In semiconductor devices, this effect arises due to random diffusion of minority carriers and the random generation and recombination of hole-electron pairs. Its power spectral density is known as, noise power spectral density.

Where,

τ_q = Transit time
q = Charge of electron
I_0 = Average value of current.

The power density spectrum of shot noise is considered as constant for frequencies below.

$$S_i(\omega) = qI_0$$

Thermal noise arises due to the random motion of free electrons in a conducting medium. The PSD of current due to free electrons is given by,

$$S_i(\omega) = \frac{2kTG\alpha^2}{\alpha^2 + \omega^2} = \frac{2kTG}{1 + \left(\frac{\omega}{\alpha}\right)^2}$$

Where,

k = Boltzmann constant
T = Ambient temperature
G = Conductance of the resistor
α = Average number of collisions per second of an electron.

(ii) **Noise suppression in semiconductor devices:** For an ideal amplifier (noiseless amplifier) noise figure is unity. The closeness of the noise figure to unity is the measure of superiority of the amplifier from the noise point of view. However, noise figure measures not the absolute but the relative quality of the amplifier. It indicates the noisiness source.

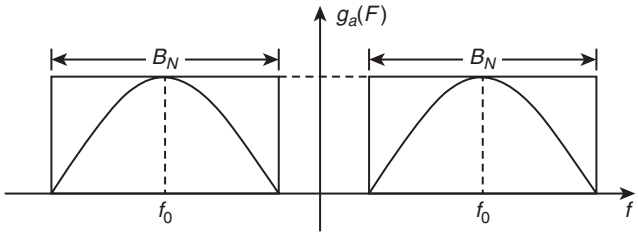
It is evident from the defirmition that the noise figure of an amplifier can be made extra noise in the source. This is not the proper solution for improving the performance of the amplifier since this approach merely makes the source so noisy that the amplifier in comparison appears almost noise free.

The overall *S*/*N* ratio at the output, however, deteriorates badly and consequently the output signal is much more noisy. It is therefore important not to increase the noise in the source in order to improve the noise figure. A step up transformer used at the input increases the input noise as well as the input signal. The increased noise at the source makes the amplifier look less noisy without deteriorating the *S*/*N* of the input. Hence, the noise figure is reduced and *S*/*N* at the output terminals actually improves.

EXAMPLE 15.11

How noise equivalent bandwidth of electronic circuit can be estimated?

Solution: Consider a system with the available gain characteristic as shown in the following figure:



Consider an ideal filter with bandwidth *BN*, whose centre frequency is also f_0. The bandwidth *BN* is adjusted such that the ideal filter and the real system pass the same amount of noise power. This *BN* is the noise bandwidth of the real circuit. Thus, the noise bandwidth of an electronic circuit is the real circuit. Assume a thermal noise source with two-sided PSD $\frac{KT}{2}$ available at the input of the circuit whose gain is available.

The output noise power is $N_0 = \frac{KT}{2}\int_{-\infty}^{\infty} g_a(f)\cdot df$

Assuming the same source connected to ideal system, output noise power $(N_0)_{\text{Ideal}}$ is equal to

$$2\times\left(\frac{KT}{2}\right)g_{a0}B_N = g_{a0}KTB_N$$

To define noise bandwidth,

$$g_aKTB_N = \frac{KT}{2}\int_{-\infty}^{\infty} g_a(f)\cdot df$$

$$\therefore B_N = \frac{1}{2g_{a0}}\int_{-\infty}^{\infty} g_a(f)\cdot df$$

EXAMPLE 15.12

Derive the equation for narrow-band noise and illustrate all its properties.

Solution: Consider a sample function of noise and select an interval of duration *T* from that sample function. If we generate a periodic waveform, in which the waveform in the selected interval is repeated for every *T*s, then this periodic waveform can be expanded using Fourier series and such series represent the noise sample function in the interval $\left(-\frac{T}{2}+\frac{T}{2}\right)$.

$$\eta_t^s(t) = \sum_{k=1}^{\infty}(ak\cos 2\pi\, k\Delta f\, T + bk\sin 2\pi\, k\,\Delta f\, t)$$

As $T\to\infty$; $\Delta f\to 0$ and the periodic sample functions of noise will become the actual noise sample function $\eta(t)$. Consider a frequency f_0 to correspond to $K = K_1$

$$\text{i.e.,}\qquad f_0 = K_1\cdot\Delta f$$

$$\text{i.e.,}\qquad 2\pi\Delta f\, t - 2\pi K_1\,\Delta f t = 0$$

$$\therefore \eta(t) = \lim_{\Delta f\to 0}\sum_{K=1}^{\infty}\left[a_k\cos 2\pi\left[f_0+(K-K_1)\Delta f\right]t + b_k\sin 2\pi\left[f_0+(K-K_1)\Delta f\right]t\right]$$

$$= \lim_{\Delta f\to 0}\sum_{K=1}^{\infty}\left[a_K\cos 2\pi(K-K_1)\Delta f t\cdot\cos 2\pi f_0 t - a_K\sin 2\pi(K-K_1)\Delta f t\cdot\sin 2\pi f_0 t + b_K\sin 2\pi f_0 t\cdot\cos 2\pi(K-K_1)\Delta f t + b_K\sin 2\pi(K-K_1)\Delta f t\cdot\cos 2\pi f_0 t\right]$$

$$= \eta_c(t)\cos 2\pi f_0 t - \eta_s(t)\sin 2\pi f_0 t$$

Where,

$$\eta_c(t) = \lim_{\Delta f \to 0}\left[\sum_{K=1}^{\infty}(a_K \cos 2\pi(K-K_1)\Delta f t + b_K \sin 2\pi(K-K_1)\Delta f t)\right]$$

$$\eta_s(t) = \lim_{\Delta f \to 0}\left[\sum_{K=1}^{\infty}(a_K \sin 2\pi(K-K_1)\Delta f t - b_K \cos 2\pi(K-K_1)\Delta f t)\right]$$

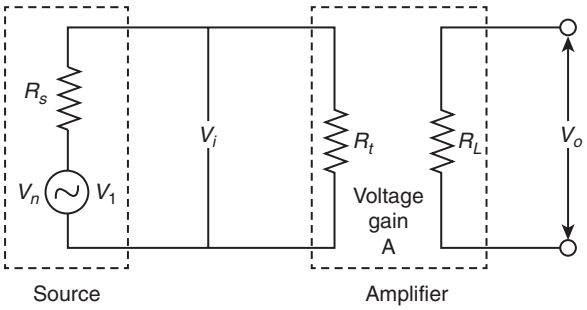
Properties:

1. $E[\eta_c(t) - \eta_s(t)] = 0$
2. $E[\eta_c^2(t)] = E[\eta_s^2(t)]$
3. $\eta_c(t)$ and $\eta_s(t)$ are of same autocorrelation function.
4. $\eta_c(t)$ and $\eta_s(t)$ are Gaussian processes.

EXAMPLE 15.13

Describe a method to determine the internal noise of and amplifier?

Solution: Internal noise is nothing but the noise created by the amplifier itself. This can be determined by modelling using thermal noise. Consider the block diagram for calculating the internal noise power.



The amplifier has an input impedance R_t and output impedance R_L and voltage gain A. H is fed from a source of internal impedance R_s. Define equivalent noise resistance that does not incorporate R_s.

$$R'_{eq} = R_{eq} - R_s$$

Total equivalent noise resistance of the amplifier is

$$R = R'_{eq} + \frac{R_s R_t}{R_s + R_t}$$

Equivalent input noise voltage is

$$V_{ni} = \sqrt{4kT + \Delta f R}$$

The noise output is $V_{no} = A V_{ni}$

$$P_{no} = \frac{V_{no}^2}{R_L} = \frac{(AV_{ni})^2}{R_L} = \frac{A^2 4kT\,\Delta f R}{R_L}$$

EXAMPLE 15.14

Derive the mathematical description of noise figure.

Solution:

Noise figure: Noise figure F of a noisy two port network is defined as the ratio of total available noise PSD at the output of a two port network to the noise PSD at the output of the two port network, assuming the network is noiseless.

$$F = \frac{G_{ao}}{G'_{ao}}$$

Generally, $G_{ao} > G'_{ao}$, because of the internal noise within the two port network. Hence, for a practical network, $F > 1$. If the two port is noiseless, $G_{ao} = G'_{ao}$ and $F = 1$. Thus, for an ideal network, $F = 1$.

Assume that the two port network is being driven by a thermal noise source. The two-sided available noise PSD is $\frac{KT}{2}$. Let $g_a(f)$ be the available gain of the network.

Therefore, $G_{ao} = g_a(f) \cdot \frac{K}{2}(T_0 + T_e)$

where T_0 is the noise temperature of the source.

Considering a noiseless two port network,

$$G'_{ao} = g_a(f) \cdot \frac{KT_0}{2}$$

Therefore, the noise figure $F = \frac{G_{ao}}{G'_{ao}} = \frac{T_0 + T_e}{T_0} = 1 + \frac{T_e}{T_0}$

Let input signal PSD be $G_{ai}^{(s)}$ and input noise PSD be G_{ai}^{n}. Let g_a be the available gain of two port network

The output signal PSD $= g_a \cdot G_{ai}^{(s)} = G_{ao}^{(s)}$

We have,

$$G_{ao} = F \cdot G'_{ao}$$

$$G_{ao}^{(n)} = F \cdot G'^{(n)}_{ao} = F \cdot g_a \cdot G_{ai}^{(n)}$$

Consider,

$$\frac{G_{ao}^{(s)}}{G_{ao}^{(n)}} = \frac{g_a \cdot G_{(ai)}^{(s)}}{F g_a \cdot G_{ai}^{(n)}}$$

$$F = \frac{G_{ai}^{(s)} \cdot G_{ao}^{(n)}}{G_{ao}^{(s)} \cdot G_{ai}^{(n)}} = \frac{\frac{G_{ai}^{(s)}}{G_{ai}^{(n)}}}{\frac{G_{ao}^{(s)}}{G_{ao}^{(n)}}}$$

In the above equation the numerator is the ratio of the input signal to noise PSD ratio and the denominator is the output signal to noise PSD ratio.

EXAMPLE 15.15

Define signal to noise ratio and noise figure and equivalent noise temperature of receiver.

Solution:

(i) **Signal to noise ratio:** Signal to noise ratio of receiver is defined as the ratio of signal power to the noise power at the same point in the receiver.

$$\text{Mathematically, SNR} = \frac{\text{Signal power}}{\text{Noise power}} = \frac{P_S}{P_N}$$

$$= \frac{\frac{V_x^2}{R}}{\frac{V_n^2}{R}} = \left(\frac{V_s}{V_n}\right)^2$$

The larger the SNR, the better the signal quality. Hence, it serves as an important measure of noise of a receiver.

(ii) **Noise figure:** Noise factor F is defined as the ratio of SNR at the input of the receiver to the SNR at the output of the receiver.

$$F = \frac{\text{input SNR}}{\text{output SNR}}$$

$$= \frac{(\text{SNR})_i}{(\text{SNR})_o} = \frac{\frac{S_i}{N_i}}{\frac{S_o}{N_o}}$$

(iii) **Equivalent noise temperature (T_{eq}):** It is defined as the temperature at which a noisy resistor has to be maintained such that by connecting the resistor to the input of a noiseless receiver it produces the same available noise power at the output of the receiver as that produced by all the sources of noise in an actual receiver.

$$T_{eq} = T_0(F-1)$$

Any system produces certain internal noise and thus results in a lower output SNR than the input SNR.

Hence, F > 1 for practical receiver
= 0 for ideal receiver.

EXAMPLE 15.16

Describe the behaviour of zero mean stationary Gaussian band limited white noise.

Solution: Many types of noise sources are Gaussian by virtue of central limit theorem and have spectral densities that are flat over a wide range of frequencies. A noise signal having a flat PSD over a wide range of frequencies is called white noise by analogy to white light. The PSD of white noise is denoted by

$$G(f) = \frac{\eta}{2} \text{ watts/Hz}$$

The factor 2 is included to indicate that $G(f)$ is a two-sided PSD. The autocorrelation function of the white noise is

$$R(\tau) = F^{-1}[G(f)] = \frac{\eta}{2} \cdot \sigma(\tau)$$

Any two different samples of a zero mean Gaussian white noise are uncorrelated and hence independent. If the input of an LTI system is a zero mean stationary Gaussian white noise $V(t)$ with PSD = $\frac{\eta}{2}$, the output of the filter will also be zero mean stationary Gaussian process with PSD of $\frac{\eta}{2}$ over a frequency range which will be decided by the system. Since the PSD of white noise is $\frac{\eta}{2}$, its average power is $\int_{-\infty}^{\infty} \frac{\eta}{2} df$ which will be infinity.

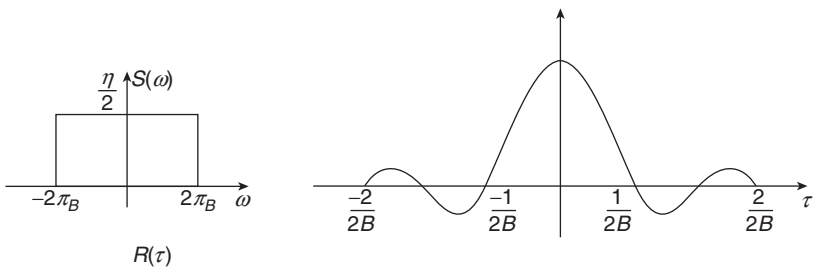
As such, white noise having infinite mean power is not physically realizable. Since the bandwidth of any real system is always finite and their average powers are also finite, we can restrict the white noise spectrum over a finite band of frequencies. Noise signals having a non-zero and constant PSD over a finite frequency band and zero elsewhere is called band-limited white noise. The spectrum for the band-limited white noise is written as

$$S(\omega) = \frac{\eta}{2} \text{ for } |\omega| \le 2\pi B$$

where B is physical bandwidth. The corresponding autocorrelation is

$$R(\tau) = \eta B\left[\frac{\sin 2\pi B\tau}{2\pi B\tau}\right]$$

These are plotted as follows:



Thus, if the process is sampled at a rate of $2B$ samples/s, the resulting noise samples are uncorrelated and being Gaussian, they are statistically independent. In generational, a random process $X(t)$ is called white noise process if its values $X(t_i)$ and $X(t_j)$ are uncorrelated for every t_i and t_j.

EXAMPLE 15.17

Derive the relation between PSDs of input and output random process of an LTI system.

Solution: Consider an LTI system whose input is a WSS process $X(t)$ and output is $Y(t)$. Let $h(\tau)$ be the impulse response of the system.

$$Y(t) = \int_{-\infty}^{\infty} h(\tau)\cdot X(t-\tau)d\tau$$

$$R_{YY}(t_1,t_2) = E\left[Y(t_1).Y(t_2)\right]$$

$$= E\left[\int_{-\infty}^{\infty} h(\tau_1)\cdot X(t_1-\tau_1)d\tau_1 \cdot \int_{-\infty}^{\infty} h(\tau_2)\,X(t_2-\tau_2)\,d\tau_2\right.$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(\tau_1)\cdot h(\tau_2)\cdot E[X(t_1-\tau_1)X(t_2-\tau_2)]d\tau_1 d\tau_2$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(\tau_1)\cdot h(\tau_2)\cdot R_{XX}(t_2-\tau_2-t_1+\tau_1)d\tau_1 d\tau_2$$

$$R_{YY}(\tau) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(\tau_1)\cdot h(\tau_2)\cdot R_{XX}(\tau-\tau_2+\tau_1)d\tau_1 d\tau_2$$

Take Fourier Transform on both the sides

$$F[R_{YY}(\tau)] = S_{YY}(\omega)$$

$$= \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(\tau_1) \cdot h(\tau_2) \cdot R_{XX}(\tau - \tau_2 + \tau_1) d\tau_1 d\tau_2 \right] e^{-j\omega\tau} d\tau$$

$$= \int_{-\infty}^{\infty} h(\tau_1) \cdot e^{j\omega\tau_1} d\tau_1 \cdot \int_{-\infty}^{\infty} h(\tau_2) e^{-j\omega\tau_2} d\tau_2 * \int_{-\infty}^{\infty} R_{XX}(\tau - \tau_2 + \tau_1) \cdot e^{-j\omega(\tau - \tau_2 + \tau_1)} d\tau$$

$$S_{YY}(\omega) = H^*(\omega) \cdot H(\omega) \cdot S_{XX}(\omega)$$

$$\therefore S_{YY}(\omega) = |H(\omega)|^2 \cdot S_{XX}(\omega)$$

DEFINITIONS AT A GLANCE

Continuous Random Processes: A continuous random process is one in which the random variable X is continuous and t is also continuous, that is, both t and s are continuous.

Discrete Random Process: A random process is said to be discrete random process if s is discrete and t is continuous.

Deterministic Random Processes: A process is called deterministic random process if future values of any sample function can be predicted from past values.

Non-deterministic Random Process: If the future values of any sample function cannot be predicted exactly from observed past values, the process is called non-deterministic random process.

Stationarity: A random process is said to be stationary if all its statistical properties do not change with time.

Autocorrelation: The autocorrelation of the random process $X(t)$ is the expected value of the product $X(t_1)\, X(t_2)$.

Wide Sense Stationary Process: A stationary random process $X(t)$ is called WSS random process (also called as weak sense stationary process) if it satisfies the following conditions:

- The mean value of the process is a constant.
- Its autocorrelation function depends only on τ and not on t.

Ergodicity: A random process $\{X(t)\}$ is said to be ergodic, if its ensemble averages are equal to appropriate time averages.

White Noise: A sample function $n(t)$ of a WSS noise random process $N(t)$ is called white noise if the power density spectrum of $N(t)$ is a constant at all frequencies.

Coloured Noise: Any noise that is not white is coloured.

FORMULAE AT A GLANCE

- **Distribution and Density Functions:** The distribution function associated with the random variable $X_1 = X(t_1)$ for a particular time t_1 is defined as

$$F_X(x_1; t_1) = P\{X(t_1) \le x_1\}$$

The second order joint distribution function for two random variables $X_1 = X(t_1)$ and $X_2 = X(t_2)$ is the two dimensional extension of the above.

$$F_X(x_1, x_2; t_1, t_2) = P\{X(t_1) \le x_1, X(t_2) \le x_2\}$$

- **First Order Stationary Process:** A random process is called first order stationary process if its first order density function does not change with a shift in time origin.

  (i.e.,) $f_X(x_1; t_1) = f_X(x_1; t_1 + \Delta)$ for any t_1 and any real number Δ

- **Cross Correlation Function:** Let $X(t)$ and $Y(t)$ be two random processes. We say that they are jointly WSS if each satisfies (4) and (5). Then their cross correlation function is given by

$$R_{XY}(t_1, t_2) = E[X(t_1)Y(t_2)]$$

- **Autocorrelation Function:** The autocorrelation of the random process $X(t)$ is the expected value of the product $X(t_1)\, X(t_2)$ and is given by

$$R_{X_1X_2}(t_1, t_2) = E\{X(t_1)X(t_2)\} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} X_1X_2 f_{X_1X_2}(X_1, X_2; t_1, t_2)dx_1dx_2$$

- **Time Averaged Mean:** The time averaged mean is defined as

$$<\mu_X> = \underset{T\to\infty}{\text{Lt}} \frac{1}{T}\int_{-\frac{T}{2}}^{\frac{T}{2}} X(t)dt$$

- **Mean-Ergodic Theorem:** If $\{X(t)\}$ is a random process with constant mean μ and if $<\mu_X> = \underset{T\to\infty}{\text{Lt}} \frac{1}{T}\int_{-\frac{T}{2}}^{\frac{T}{2}} X(t)dt$ then $\{X(t)\}$ is mean-ergodic provided $\underset{T\to\infty}{\text{Lt}}[\text{Var} <\mu_X>] = 0$

- **Correlation Ergodic Process:** The stationary process $\{X(t)\}$ is correlation ergodic if

$$<\mu_2> = \underset{T\to\infty}{\text{Lt}} \frac{1}{T}\int_{-\frac{T}{2}}^{\frac{T}{2}} X(t+\lambda)X(t)$$
$$= E\{X(t+\lambda)X(t)\}$$

- **The Stationary Process:** The stationary process $\{X(t)\}$ is distribution ergodic, if

$$<\mu_Y> = \underset{T\to\infty}{\text{Lt}} \frac{1}{T}\int_{-\frac{T}{2}}^{\frac{T}{2}} Y(t)dt$$

- **Power Density Spectrum:** The spectral properties of a deterministic signal $x(t)$ are contained in its Fourier transform $X(\omega)$ which is given by

$$X(\omega) = \int_{-\infty}^{\infty} x(t)e^{-i\omega t}dt$$

  The function $X(\omega)$ is called the spectrum of $x(t)$.

OBJECTIVE TYPE QUESTIONS

1. A random process in which the random variable X and t are continuous is called_____.
   (a) Discrete random process (b) Continuous random process
   (c) Deterministic random process (d) none

2. A random process in which if future values of any sample function can be predicted from past values is called _____.
   (a) Discrete random process (b) Continuous random process
   (c) Deterministic random process (d) none

3. A random process in which all the statistical properties do not change with time is called _____.
   (a) Deterministic random process (b) Non-deterministic random process
   (c) Stationary random process (d) none

4. A first order stationary random process has a _____.
   (a) Constant derivative (b) Constant mean
   (c) Constant variance (d) none

5. The auto correlation function of the random process $X(t)$ is given by $R_{XX}(t_1, t_2) =$ _____.
   (a) $E[X(t_1)X(t_2)]$ (b) $E[X(t_1)+X(t_2)]$
   (c) $E[X(t_1)-X(t_2)]$ (d) none

6. If the ensemble averages are equal to appropriate time averages then random process is called _____.
   (a) Stationary processes (b) mean Ergodic processes
   (c) Evolutionary processes (d) none

7. For wide sense stationary processes, we have $R_{XX}(0) =$ _____.
   (a) $E(X(t))$ (b) $E(X(t)\,X(t+\tau))$
   (c) $E[X^2(t)]$ (d) none

8. From the properties of cross correlation function, one of the following is true.
   (a) $|R_{XY}(\tau)| \le \sqrt{R_{XX}(0)R_{YY}(0)}$ (b) $|R_{XY}(\tau)| \le \frac{1}{2}[R_{XX}(0)+R_{YY}(0)]$
   (c) Both (a) and (b) (d) none

9. A WSS noise random process in which the power density spectrum is constant at all frequencies is called _____.
   (a) White noise (b) Coloured noise
   (c) Band-limited noise (d) none

10. If T_s is the sampling interval, the finite band of a band-limited process, ω_s is given by
   (a) $\frac{T_s}{2\pi}$ (b) $\frac{2\pi}{T_s}$
   (c) $2\pi\, T_s$ (d) none

ANSWERS

1. (b) 2. (c) 3. (c) 4. (b) 5. (a) 6. (b) 7. (c) 8. (c)
9. (a) 10. (b)

16 Advanced Random Process

Prerequisites

Before you start reading this unit, you should:

- Be able to distinguish different random processes like continuous random process, discrete random process, deterministic random process, non-deterministic random process
- Know stationary random process, WSS, cross correlation function, Ergodic random process
- Know power density spectrum and its properties
- Know discrete time processes and sequences

Learning Objectives

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

- Know the mean and auto correlation of the Poisson process
- Know about Markov process, Chapman-Kolmogorov theorem
- Find the applications of queues
- Know about Gaussian process and Narrow band Gaussian process

INTRODUCTION

Sometimes we may be interested to study the times at which components fail in a large system or the times of arrival of phone calls at an exchange. Here, we are not just interested in the event, but in the sequence of random time instants at which the events occur.

16.1 POISSON PROCESS

Definition

If $X(t)$ represents the number of occurrences of a certain event in $(0, t)$, then the discrete random process $X(t)$ is called the Poisson process if the following conditions are satisfied:

(i) $P[1 \text{ occurrence in } (t, t + \Delta t)] = \lambda \Delta t + 0(\Delta t)$

(ii) $P[0 \text{ occurrence in } (t, t + \Delta t)] = 1 - \lambda \Delta t + 0(\Delta t)$

(iii) $P[2 \text{ or more occurrences in } (t, t + \Delta t)] = 0(\Delta t)$

(iv) $X(t)$ is independent of the number of occurrences of the event in any interval prior and after the interval $(0, t)$.

(v) The probability that the event occurs a specified number of times in $(t_0, t_0 + t)$ depends only on t but not on t_0.

Probability Law for the Poisson Process $X(t)$

Let λ be the number of occurrences of the event in unit time.

Let $P_n(t) = P\{X(t) = n\}$

$P_n(t + \Delta t) = P\{X(t + \Delta t) = n\}$

$= P\{(n-1)$ calls in $(0, t)$ and 1 call in $(t, t + \Delta t)\} + P\{n$ calls in $(0, t)$ and no call in $(t, t + \Delta t)\}$

$= P_{n-1}(t)\lambda\Delta t + P_n(t)(1 - \lambda\Delta t)$ {from the above conditions}

$$\therefore \frac{P_n(t+\Delta t) - P_n(t)}{\Delta t} = \lambda\{P_{n-1}(t) - P_n(t)\}$$

$$\underset{\Delta t \to 0}{\text{Lt}} \frac{P_n(t+\Delta t) - P_n(t)}{\Delta t} = \underset{\Delta t \to 0}{\text{Lt}} \lambda\{P_{n-1}(t) - P_n(t)\}$$

$$\frac{dP_n(t)}{dt} = \lambda\{P_{n-1}(t) - P_n(t)\} \tag{16.1}$$

Let the solution of equation (16.1) be

$$P_n(t) = \frac{(\lambda t)^n}{n!} f(t) \tag{16.2}$$

Differentiating equation (16.2) w.r.t. t,

$$P_n'(t) = \frac{\lambda^n}{n!}\left[nt^{n-1} f(t) + t^n f'(t)\right] \tag{16.3}$$

From equations (16.1), (16.2) and (16.3)

$$\frac{\lambda^n}{n!} t^n f'(t) = -\lambda \frac{(\lambda t)^n}{n!} f(t)$$

$$f'(t) = -\lambda f(t)$$

$$f(t) = Ke^{-\lambda t}$$

From equations (16.2) $f(0) = P_0(0) = P\{X(0) = 0\}$

$= P\{$no event occurs in $(0, 0)\}$

$= 1$

$$\therefore f(t) = e^{-\lambda t}$$

$$\therefore P_n(t) = P\{X(t) = n\} = \frac{e^{-\lambda t}(\lambda t)^n}{n!}, \quad n = 0, 1, 2, \ldots$$

Thus, the probability distribution of $X(t)$ is the Poisson distribution with parameters λt.

Second Order Probability Function of a Homogeneous Poisson Process

In the previous topic, we have seen probability function for a first order. Now an extension to it is observed here.

$$P[X(t_1) = n_1, X(t_2) = n_2]$$

$$= P[X(t_1) = n_1]\, P\left[\frac{X(t_2) = n_2}{X(t_1) = n_1}\right], \quad t_2 > t_1$$

$$= P[X(t_1) = n_1]\ P[\text{The event occurs } (n_2 - n_1) \text{ times in the interval } (t_2 - t_1)]$$

$$= \frac{e^{-\lambda t_1}(\lambda t_1)^{n_1}}{(n_1)!} \frac{e^{-\lambda(t_2 - t_1)}\{\lambda(t_2 - t_1)\}^{n_2 - n_1}}{(n_2 - n_1)!}, \quad n_2 \geq n_1$$

$$= \begin{cases} \dfrac{e^{-\lambda t_2} \lambda t_1^{n_1} (t_2 - t_1)^{(n_2 - n_1)}}{(n_1)!(n_2 - n_1)!}, & n_2 \geq n_1 \\ 0, & \text{otherwise} \end{cases}$$

In a similar way we can get probability function of third order also.

$$P\{X(t_1) = n_1,\ X(t_2) = n_2,\ X(t_3) = n_3\}$$

$$= \begin{cases} \dfrac{e^{-\lambda t_3} \lambda^{n_3} t_1^{n_1} (t_2 - t_1)^{n_2 - n_1} (t_3 - t_2)^{n_3 - n_2}}{(n_1)!\,(n_2 - n_1)!\,(n_3 - n_2)!}, & n_3 \geq n_2 \geq n_1 \\ 0, & \text{elsewhere} \end{cases}$$

16.2 MEAN AND AUTO CORRELATION OF THE POISSON PROCESS

The probability law of the Poisson process $\{X(t)\}$ is the same as that of a Poisson distribution with parameter λt.

$$E[X(t)] = \text{var}\,[X(t)] = \lambda t \tag{16.4}$$

$$E[X^2(t)] = \lambda t + \lambda^2 t^2 \tag{16.5}$$

Auto correlation function, $R_{XX}(t_1, t_2) = E\,\{X(t_1)X(t_2)\}$

$$= E\{X(t_1)\}\{X(t_2) - X(t_1) + X(t_1)\}$$

$$= E[X(t_1)\{X(t_2) - X(t_1)\}] + E\{X^2(t_1)\}$$

$$= E[X(t_1)E[X(t_2) - X(t_1)] + E\{X^2(t_1)\}$$

Since $X(t)$ is a process of independent increments,

$$R_{XX}(t_1, t_2) = \lambda t_1[\lambda(t_2 - t_1)] + \lambda(t_1) + \lambda t_1^2,\ t_2 \geq t_1$$

$$= \lambda^2 t_1 t_2 + \lambda t_1 \quad \text{if } t_2 \geq t_1 \quad \text{[From (16.4) and (16.5)]}$$

$$R_{XX}(t_1, t_2) = \lambda^2 t_1 t_2 + \lambda \min(t_1, t_2)$$

Covariance function, $C_{XX}(t_1, t_2) = R_{XX}(t_1, t_2) - E\{X(t_1)\}E\{X(t_2)\}$

$$\therefore C_{XX}(t_1, t_2) = \lambda^2 t_1 t_2 + \lambda t_1 - \lambda^2 t_1 t_2$$

$$= \lambda t_1 \qquad \text{if } t_2 \geq t_1$$

$$= \min(t_1, t_2)$$

Hence the correlation coefficient function is given by

$$r_{XX}(t_1, t_2) = \frac{C_{XX}(t_1, t_2)}{\sqrt{\text{var}[X(t_1)]}\sqrt{\text{var}[X(t_2)]}}$$

$$= \frac{\lambda t_1}{\sqrt{\lambda t_1 \lambda t_2}} = \frac{t_1}{t_2}, \quad t_2 > t_1$$

Properties

1. The Poisson process is a Markov process, that is, the conditional probability distribution of $X(t_3)$ given all the past values $X(t_1) = n_1$, $X(t_2) = n_2$ depends only on the most recent values $X(t_2) = n_2$, which is the Markov property.
2. Sum of two independent Poisson processes is a Poisson process, that is, if $X(t_1)$, $X(t_2)$ are two independent Poisson processes then $X(t_1) + X(t_2)$ is a Poisson process with parameter $(\lambda_1 + \lambda_2)t_1$.
3. Difference of two independent Poisson processes is not a Poisson process, that is, If $X(t_1)$ and $X(t_2)$ are two independent Poisson processes then $[X(t_1) - X(t_2)]$ is not a Poisson process.
4. The inter-arrival time of a Poisson process, that is, the interval between two successive occurrences of a Poisson process with parameter λ has an exponential distribution with mean $\frac{1}{\lambda}$.

16.3 MARKOV PROCESS

Markov processes represent the simplest generalization of independent processes by permitting the outcome at any instant to depend only on the outcome that precedes it and none before that. Thus, in Markov process $X(t)$, the past has no influence on the future if the present is specified.

A random process $X(t)$ is called a Markov process if

$$P\left\{\frac{X(t_n) = a_n}{X(t_{n-1}) = a_{n-1}}, X(t_{n-2}) = a_{n-2}, \ldots, X(t_2) = a_2, X(t_1) = a_1\right\}$$

$$= P\left\{\frac{X(t_n) = a_n}{X(t_{n-1}) = a_{n-1}}\right\}$$

for all $t_1 < t_2 < \ldots < t_n$

If the future behaviour of a process depends only on the present value, but not on the past, then the process is called a Markov process.

If the above condition is satisfied for all n, then the process $X(t)$ is called a Markov chain and the constants $\{a_1, a_2, \ldots, a_n\}$ are called the states of the Markov chain.

Consider the problem of tossing a coin 'n' times. The total number of heads in n tosses will be equal to K, $K = 0, 1, 2, \ldots, n$. If the random variable S_{n+1} represents the total number of heads in $(n + 1)$ tosses; then it can take values of either $(K + 1)$ [if the $(n + 1)^{\text{th}}$ toss yields a head] or K [if the $(n + 1)^{\text{th}}$ toss gives a tail].

$$\text{i.e., } P\left[\frac{S_{n+1} = (K+1)}{S_n = K}\right] = \frac{1}{2} \text{ and}$$

$$P\left[\frac{S_{n+1} = K}{S_n = K}\right] = \frac{1}{2}$$

Thus, the value of the random variable S_{n+1} depends on the value of S_n alone and not on the other previous values of $S_1, S_2, \ldots, S_{n-1}$. Such a random process is called Markov process. Here, $\{a_1, a_2, \ldots, a_n, \ldots\}$ are called the states of the Markov chain. The conditional probability $P\left\{\dfrac{X_n = a_j}{X_{n-1} = a_i}\right\}$ is called one-step transition probability, from state a_i to state a_j at the n^{th} step. It is denoted by $P_{ij}(n-1, n)$. In addition, the matrix $P = \{P_{ij}\}$ is known as one-step transition probability matrix (tpm).

Now, the conditional probability that the process is in state a_j at step n, given that it was in state a_i at step 0, i.e., $P\left\{\dfrac{X_n = a_j}{X_0 = a_i}\right\}$ is called the n-step transition probability.

It is denoted by $p_{ij}^{(n)}$.

If the probability that the process is in state a_i is $p_i (i = 1, 2, \ldots, K)$ at any arbitrary step, then the row vector $p = (p_1, p_2, \ldots, p_K)$ is called the probability distribution of the process at that time. In particular $P^{(0)} = \{p_1^{(0)}, p_2^{(0)}, \ldots, p_k^{(0)}\}$ is called the initial probability distribution.

Hence, a Markov chain $\{X_n\}$ is completely determined by the transition probability matrix together with initial probability distribution.

16.4 CHAPMAN-KOLMOGOROV THEOREM

Statement: If P is the tpm of a homogeneous Markov chain, then the n-step tpm $P^{(n)}$ is equal to P_n. (i.e.,) $P_n = [P_{ij}^{(n)}] = [P_{ij}]^n$.

Proof: We know that

$P_{ij}^{(2)} = P\left\{\dfrac{X_2 = j}{X_0 = i}\right\}$, Since the chain is homogeneous.

The state j can be reached from state i in two steps through some intermediate state K.

$$\begin{aligned}
\text{Now, } P_{ij}^{(2)} &= P\left\{\frac{X_2 = j}{X_0 = i}\right\} \\
&= P\left\{X_2 = j, \frac{X_1 = K}{X_0 = i}\right\} \\
&= P\left\{\frac{X_2 = j}{X_1 = K}, X_0 = i\right\} \cdot P\left\{\frac{X_1 = K}{X_0 = i}\right\} \\
&= P_{Kj}^{(1)} \cdot P_{iK}^{(1)} \\
&= P_{iK} P_{Kj}
\end{aligned}$$

Since the transition from state i to state j in two steps can take place through any one of the intermediate states, k can assume the values 1, 2, 3, … The transitions through various intermediate states are mutually exclusive. Hence,

$$P_{ij}^{(2)} = \sum_k P_{ik} P_{kj}$$

We can observe that the tpm of ij^{th} element of second stage is equal to product of two one-step tpms.

In other words, $P^{(2)} = P^2$

Now consider,

$$P_{ij}^{(3)} = P\left\{\frac{X_3 = j}{X_0 = 1}\right\}$$

$$= \sum_k P\left\{\frac{X_3 = j}{X_2 = k}\right\} \cdot P\left\{\frac{X_2 = k}{X_0 = i}\right\}$$

$$= \sum_k P_{kj} P_{ik}^{(2)}$$

$$= \sum_k P_{ik}^{(2)} P_{kj} = P^2 \cdot P = P^3$$

Similarly, $P_{ij}^{(3)} = \sum_k P_{ik} P_{ik}^{(2)} = P^2 \cdot P = P^3$

$$\therefore P^{(3)} = P^2 \cdot P = P \cdot P^2 = P^3$$

Proceeding in a similar way, we get

$$P^{(n)} = P^n$$

16.5 DEFINITIONS IN MARKOV CHAIN

A stochastic matrix P is said to be a regular matrix if all the entries of P^m are positive. A homogeneous Markov chain is said to be regular if its tpm is regular.

Irreducible Markov Chain

If $P_{ij}^{(n)} > 0$ for some n and for all i and j then every state can be reached from every other state. When this condition is satisfied the Markov chain is said to be irreducible. The tpm of an irreducible chain is an irreducible matrix otherwise the chain is said to be non-Irreducible or reducible.

Period

The state i of a Markov chain is called a return state if $P_{ij}^{(n)} > 0$ for some $n \geq 1$. The period d_i of a return state i is defined as the greatest common divisor of all m such that $P_{ii}^{(n)} > 0$.

$$d_i = \text{GCD } \{m: P_{ij}^{(m)} > 0\}$$

Hence, state i is said to be periodic with period d_i if $d_i > 1$ and aperiodic if $d_i = 1$ & $p_{ii} \neq 0$.

The probability that the chain returns to state i, having started from state i for the first time at the n^{th} step(or after n transitions) is denoted by $f_{ii}^{(n)}$ and called the first return time probability or the recurrence time probability.

If $F_n = \sum_{n=1}^{\infty} f_{ii}^{(n)} = 1$, the return to state '$i$' is certain.

The mean recurrence time of the state i is given by $\mu_{ii} = \sum_{n=1}^{\infty} n f_{ii}^{(n)}$.

A state i is said to be persistent or recurrent if the return to state i is certain, that is, if $F_{ii} = 1$.

It is said to be transient if the return to state i is uncertain, that is, if $F_{ii} < 1$.

The state i is said to be non-null persistent if its mean recurrence time μ_{ii} is finite and null persistent if $\mu_{ii} = \infty$.

A non-null persistent and aperiodic state is called ERGODIC.

16.6 APPLICATION TO THE THEORY OF QUEUES

In any type of queue or a waiting line, customers (jobs) arrive randomly and wait for service. A customer receives immediate service when the server is free, otherwise the customer joins the queue and waits for service.

The server continues service according to some schedule such as first in, first out, as long as there are customers in the queue waiting for service.

The total duration of uninterrupted service beginning at $t = 0$ is known as the busy period.

Under the assumption that various arrivals and service times are mutually independent random variables, the total number of customers during the busy period, the duration of the busy period, and the probability of its termination are of special interest.

Queues are characterized depending on the type of inter-arrival time distribution as well as the service time distribution, the type of inter-arrival time distribution, and the total number of servers employed.

Let X_n denote the number of customers (jobs) waiting in line for service at the instant t_n when the n^{th} customer departs after completing service. If we consider the first customer arriving at an empty counter and receiving immediate service as representing the zeroth generation, then the direct descendants are X_1 customers arriving during the service time of the first customer and forming a waiting line. This process continues as long as the queue lasts.

16.7 RANDOM WALK

Consider an example in which we explain how the tpm is formed for a Markov chain.

Assume that a man is at an integral point of the x-axis between the origin and the point $x = 3$. He takes a unit step either to the right with probability 0.7 or to the left with probability 0.3, unless he is at the origin when he takes a step to the right to reach $x = 1$ or he is at the point $x = 3$, when he takes a step to the left to reach $x = 2$. The chain is called Random walk with reflecting barriers. The tpm is as follows:

$$\begin{array}{c} \\ \text{States of } X_{n-1} \end{array} \begin{array}{c} \text{States of } X_n \\ \begin{array}{cccc} 0 & 1 & 2 & 3 \end{array} \\ \begin{array}{c} 0 \\ 1 \\ 2 \\ 3 \end{array} \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0.3 & 0 & 0.7 & 0 \\ 0 & 0.3 & 0 & 0.7 \\ 0 & 0 & 1 & 0 \end{bmatrix} \end{array}$$

P_{21} = the element in 2nd row, 1st column of the tpm = 0.3, that is, if the process is at state 2 at step $(n - 1)$, the probability that it moves to state 1 at step $n = 0.3$, where n is any positive integer.

Worked Out Examples

EXAMPLE 16.1

The transition probability matrix of a Markov chain $\{X_n\}$ $n = 1, 2, 3, \ldots$ having 3 states 1, 2, and 3 is

$$P = \begin{bmatrix} 0.1 & 0.6 & 0.3 \\ 0.5 & 0.2 & 0.4 \\ 0.4 & 0.2 & 0.3 \end{bmatrix}$$

and the initial distribution is $P^{(0)} = (0.7, 0.2, 0.1)$. Find

(i) $P(X_2 = 3)$

(ii) $P(X_3 = 2, X_2 = 3, X_1 = 3, X_0 = 2)$.

Solution:

$$P^{(2)} = P^2 = \begin{bmatrix} 0.1 & 0.6 & 0.3 \\ 0.5 & 0.2 & 0.4 \\ 0.4 & 0.2 & 0.3 \end{bmatrix} \begin{bmatrix} 0.1 & 0.6 & 0.3 \\ 0.5 & 0.2 & 0.4 \\ 0.4 & 0.2 & 0.3 \end{bmatrix}$$

$$= \begin{bmatrix} 0.43 & 0.24 & 0.36 \\ 0.31 & 0.42 & 0.35 \\ 0.26 & 0.34 & 0.29 \end{bmatrix}$$

(i) $P(X_2 = 3) = \sum_{i=1}^{3} P\left(\frac{X_2 = 3}{X_0 = i}\right) \cdot P(X_0 = i)$

$= P_{13}^{(2)} P(X_0 = 1) + P_{23}^{(2)} P(X_0 = 2) + P_{33}^{(2)} \cdot P_{33}(X_0 = 3)$

$= 0.36 \times 0.7 + 0.35 \times 0.2 + 0.29 \times 0.1$

$= 0.351$

(ii) $P\left\{\frac{X_1 = 3}{X_0 = 2}\right\} = P_{23} = 0.4$

$P\{X_1 = 3, X_0 = 2\} = P\left\{\frac{X_1 = 3}{X_0 = 2}\right\} \times P(X_0 = 2)$

$= 0.4 \times 0.4 = 0.16$

$P\{X_2 = 3, X_1 = 3, X_0 = 2\} = P\left\{\frac{X_2 = 3}{X_4 = 3}, X_0 = 2\right\} \times P\{X_1 = 3, X_0 = 2\}$

$= P\left\{\frac{X_2 = 3}{X_1 = 3}\right\} \times P\{X_1 = 3, X_0 = 2\}$ by Markov property

$= 0.3(0.16) = 0.048$

$P\{X_3 = 2, X_2 = 3, X_1 = 3, X_0 = 2\}$

$= P\left\{\frac{X_3 = 2}{X_2 = 3}, X_1 = 3, X_0 = 2\right\} \times P\{X_2 = 3, X_1 = 3, X_0 = 2\}$

$= P\left\{\frac{X_3 = 2}{X_2 = 3}\right\} \times P\{X_2 = 3, X_1 = 3, X_0 = 2\}$ by Markov property

$= (0.2)(0.048)$

$= 0.0096$

EXAMPLE 16.2

A gambler has ₹2. He bets ₹1 at a time and wins ₹1 with probability $\frac{1}{2}$. He stops playing if he loses ₹2 or wins ₹4.

(i) What is the tpm of the related Markov chain?

(ii) What is the probability that he has lost his money at the end of 5 plays?

(iii) What is the probability that the game lasts more than 7 plays?

Solution: Let X_n represent the amount with the player at the end of n^{th} round of the play.

State space of $X_n = (0, 1, 2, 3, 4, 5, 6)$ as the game ends if the player loses all the money ($X_n = 0$) or wins ₹4 i.e., he has ₹6 ($X_n = 6$).

tpm of the Markov chain is

$$P = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{array} \begin{array}{c} \begin{array}{ccccccc} 0 & 1 & 2 & 3 & 4 & 5 & 6 \end{array} \\ \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \end{array}$$

The initial probability distribution of X_n is $P(0) = (0, 0, 1, 0, 0, 0, 0)$ as the player has got ₹ 2 to start with the game.

$$P^{(1)} = P^{(0)}P = \left(0, \frac{1}{2}, 0, \frac{1}{2}, 0, 0, 0\right)$$

$$P^{(2)} = P^{(1)}P = \left(\frac{1}{4}, 0, \frac{1}{2}, 0, \frac{1}{4}, 0, 0\right)$$

$$P^{(3)} = P^{(2)}P = \left(\frac{1}{4}, \frac{1}{4}, 0, \frac{3}{8}, 0, \frac{1}{8}, 0\right)$$

$$P^{(4)} = P^{(3)}P = \left(\frac{3}{8}, 0, \frac{5}{16}, 0, \frac{1}{4}, 0, \frac{1}{16}\right)$$

$$P^{(5)} = P^{(4)}P = \left(\frac{3}{8}, \frac{5}{32}, 0, \frac{9}{32}, 0, \frac{1}{8}, \frac{1}{16}\right)$$

$P\{$The man has lost his money at the end of 5 plays$\}$

$= P\{X_5 = 0\} =$ element corresponding to state 0 in $P^{(5)}$

$= \frac{3}{8}$

$$P^{(6)} = P^{(5)}P = \left(\frac{29}{34}, 0, \frac{7}{32}, 0, \frac{13}{64}, 0, \frac{1}{8}\right)$$

$$P^{(7)} = P^{(6)}P = \left(\frac{29}{64}, \frac{7}{64}, 0, \frac{27}{128}, 0, \frac{13}{128}, \frac{1}{8}\right)$$

P[game lasts more than 7 rounds]

$= P\{$System is neither in state 0 nor in 6 at the end of seventh round$\}$

$= P\{X_7 = 1, 2, 3, 4, \text{ or } 5\}$

$$= \left(\frac{7}{64} + 0 + \frac{27}{128} + 0 + \frac{13}{128}\right)$$

$$= \frac{27}{64}$$

16.8 GAUSSIAN PROCESS

Most of the random processes like Weiner process can be approximated by a Gaussian process according to central limit theorem.

Definition

A real valued random process $X(t)$ is called a Gaussian process or normal process if the random variables $X(t_1), X(t_2), \ldots, X(t_n)$ are jointly normal for every $n = 1, 2, \ldots$, and for any set of t_i's.

The n^{th} order density of a Gaussian process is given by

$$f(x_1, x_2, \ldots, x_n; t_1, t_2, \ldots, t_n) = \frac{1}{(2\pi)^{\frac{n}{2}} \left|[C_X]\right|^{\frac{1}{2}}} \exp\left[\frac{-1}{2[C_X]}[X - \bar{X}]^T [x - \bar{X}]\right] \tag{16.6}$$

where we define matrices

$$[\bar{X} - \bar{X}] = \begin{bmatrix} x_1 - \bar{X}_1 \\ x_2 - \bar{X}_2 \\ \cdot \\ \cdot \\ \cdot \\ x_n - \bar{X}_n \end{bmatrix}$$

$$[C_X] = \begin{bmatrix} C_{11}\ C_{12} \ldots C_{1n} \\ C_{21}\ C_{22} \ldots C_{2n} \\ \cdot \\ \cdot \\ \cdot \\ C_{n1}\ C_{n2} \ldots C_{nn} \end{bmatrix}$$

$[X - \bar{X}]^T$ denotes the transpose of the matrix $[X - \bar{X}]$

$[C_X]$ is called covariance matrix of n random variables given by

$$C_{ij} = E[(X_i - \bar{X}_i)(X_j - \bar{X}_j)]$$

$$= \sigma^2_{X_i},\ i = j$$

$$C_{X_iX_j},\ i \neq j$$

The density (16.6) is called N-variate Gaussian density function. When $n = 2$, the covariance matrix becomes

$$[C_X] = \begin{bmatrix} \sigma^2_{X_1} & \rho\sigma_{X_1}\sigma_{X_2} \\ \rho\sigma_{X_1}\sigma_{X_2} & \sigma^2_{X_2} \end{bmatrix}$$

$$[C_X]^{-1} = \frac{1}{(1-\rho^2)} \begin{bmatrix} 1/\sigma^2_{X_1} & -\rho/\sigma_{X_1}\sigma_{X_2} \\ -\rho/\sigma_{X_1}\sigma_{X_2} & 1/\sigma^2_{X_2} \end{bmatrix}$$

$$\left|[C_X]^{-1}\right| = \frac{1}{\sigma^2_{X_1}\sigma^2_{X_2}(1-\rho^2)}$$

Properties of Gaussian Processes

1. If a Gaussian process is wide-sense stationary, it is also strict-sense stationary.
2. If the input $X(t)$ of a linear system is a Gaussian process, the output will also be a Gaussian process.
3. If the membership functions of a Gaussian process are uncorrelated, then they are independent.

16.9 BAND PASS PROCESS

A process $X(t)$ is called band pass if its spectrum is 0 outside an interval (ω_1, ω_2).

$$S_{XX}(\omega) \neq 0 \text{ in } |\omega - \omega_0| \leq \frac{\omega_B}{2} \text{ and } |\omega + \omega_0| < \frac{\omega_B}{2}$$

$$= 0 \text{ in } |\omega - \omega_0| > \frac{\omega_B}{2} \text{ and } |\omega + \omega_0| > \frac{\omega_B}{2}$$

It is called narrowband or quasi-monochromatic if its band width $\omega_2 - \omega_1$ is small compared with the centre frequency.

If the power spectrum $S_{XX}(\omega)$ of a band pass process $X(t)$ is an impulse function then the process is called monochromatic.

16.10 NARROW BAND GAUSSIAN PROCESS

In communication system, information-bearing signals are often narrow band Gaussian process. When such signals are viewed on an oscilloscope, they appear like a sine wave with slowly varying amplitude and phase.

Hence a narrow band Gaussian process $X(t)$ is

$$X(t) = R_X(t)\cos(\omega_0 \pm \theta_X(t))$$

Here, $R_X(t)$ is called envelope of the low pass process $X(t)$ and θ_X is called its phase.

$$\therefore X(t) = R_X(t)\ [\cos\theta_X(t)\cos\omega_0 t \mp \sin\theta_X(t)\sin\omega_0(t)]$$

Caution:

(i) In the above low pass process, $R_X(t)\cos\theta_X(t)$ is called inphase component of the process $X(t)$ and it is denoted by $X_c(t)$ or $I(t)$.

(ii) $R_X(t)\sin\theta_X(t)$ is called quadrature component of $X(t)$ and denoted as $X_s(t)$ or $Q(t)$.

(iii) The quadrature representation of $X(t)$ is not unique and is useful only when $X(t)$ is a zero mean WSS band pass process.

Property of Narrow Band Gaussian Process

The envelope of a narrow-band Gaussian process follows a Rayleigh distribution and the phase follows a uniform distribution in $(0, 2\pi)$.

We note that $\sqrt{X_c^2(t)+X_s^2(t)} = R_X(t)$

$$\tan^{-1}\left\{\frac{X_s(t)}{X_c(t)}\right\} = \theta_X(t)$$

16.11 BAND LIMITED PROCESS

If the power spectrum of a band pass random process is zero outside some frequency band of width ω that does not include $\omega = 0$, the process is called band-limited. The concept of a band-limited process forms a convenient approximation that often allows analytical problem solutions that otherwise might not be possible.

DEFINITIONS AT A GLANCE

Poisson process: $X(t)$ is poisson process if

(i) $P(1 \text{ occurrence in } (t, t+\Delta t)) = \lambda\Delta t + 0(\Delta t)$

(ii) $P(0 \text{ occurrence in } (t, t+\Delta t)) = 1 - \lambda\Delta t + 0(\Delta t)$

(iii) $P(2 \text{ or more occurrences}) = 0(\Delta t)$

Markov process: If the future behaviour of a process depends only on the present value, but not on the past, then the process is Markov.

Regular matrix: A stochastic matrix, P is regular if all the entries of P^m are positive.

Markov chain: If $P\left\{\frac{X(t_n)=a_n}{X(t_{n-1})=a_{n-1}}, \ldots X(t_1)=a_1\right\}$ for all $t_1 < t_2 < \ldots t_n$ then $X(t)$ is called a Markov chain.

Transition probability matrix (tpm): The conditional probability $P\left\{\frac{X_n=a_j}{X_{n-1}=a_i}\right\}$ from state a_i to state a_j at the n^{th} step which is denoted by $P_{ij}(n-1, n)$ and the matrix $P = \{P_{ij}\}$ is called one-step transition probability matrix.

Irreducible Markov chain: If $P_{ij}^{(n)} > 0$ for some n and for all i and j then every state can be reached from every other state, then Markov chain is irreducible.

Gaussian Process: A real valued random process $X(t)$ is called Gaussian process if the random process $X(t_1), X(t_2) \ldots X(t_n)$ are jointly normal for every $n = 1, 2, \ldots$ and for any t_i's.

Band pass process: A process $X(t)$ is called band pass if its spectrum is 0 outside an interval (ω_1, ω_2).

Narrow band process: If the band width $\omega_2 - \omega_1$ is small compared with the centre frequency then the process is called narrow band process.

Monochromatic band process: If the power spectrum of a band process is an impulse function, then the process is called monochromatic.

FORMULAE AT A GLANCE

- Probability law of Poisson process

$$P_n(t) = P\{X(t) = n\}$$

$$P_n(t) = \frac{(\lambda t)^n}{n!} f(t)$$

$$= \frac{e^{-\lambda t}(\lambda t)^n}{n!}, \quad n = 0,1,2 \ldots.$$

- Second order probability function of a homogeneous Poisson process:

$$P[X(t_1) = n_1,\ X(t_2) = n_2]$$

$$= \frac{e^{-\lambda t_2} \lambda t_1{}^{n_1} (t_2 - t_1)^{(n_2 - n_1)}}{(n_1)!\,(n_2 - n_1)!},\ n_2 \geq n_1$$

- Mean of the Poisson process $E[X(t)] = \lambda t$
- Auto correlation function of the Poisson process is

$$R_{XX}(t_1, t_2) = \lambda^2 t_1 t_2 + \lambda \min(t_1, t_2)$$

- Correlation coefficient of the Poisson process is

$$R_{XX}(t_1, t_2) = \frac{t_1}{t_2},\ t_2 > t_1$$

- Chapman-Kolmogorov theorem: If P is the tpm of a homogeneous Markov chain, then the n-step tpm $P^{(n)}$ is equal to P_n (i.e.,)

$$P_n = [P_{ij}^{(n)}] = [P_{ij}]^n$$

- The n^{th} order density of a Gaussian process is

$$f(x_1, x_2 \ldots x_n; t_1, t_2, \ldots tn)$$

$$= \frac{1}{(2\pi)^{n/2} |[C_X]|^{1/2}} \exp\left[\frac{-1}{2[C_X]}[X - \bar{X}]^T [X - \bar{X}]\right]$$

- The envelope of a narrow band Gaussian process follows a Rayleigh distribution and the phase follows a uniform distribution in $(0, 2\pi)$ (i.e.,)

$$\sqrt{X_C{}^2(t) + X_s{}^2(t)} = R_X(t) \text{ and}$$

$$\tan^{-1}\left\{\frac{X_s(t)}{X_C(t)}\right\} = \theta_X(t)$$

OBJECTIVE TYPE QUESTIONS

1. The probability distribution of Poisson distribution of $X(t)$ is given by $P_n(t) =$ _____

   (a) $e^{-\lambda t}(\lambda t)^n$ (b) $\frac{e^{-\lambda t}}{n!}$

   (c) $\frac{e^{-\lambda t}(\lambda t)^n}{n!}$ (d) none

2. The second order probability function of a homogeneous Poisson process $P[X(t_1) = n_1, X(t_2) = n_2]$ is

   (a) $\frac{e^{-\lambda t_2} e^{-\lambda t_1}}{n_1!(n_2 - n_1)!}$ (b) $\frac{e^{-\lambda t_2} \lambda {t_1}^{n_1} (t_1 - t_2)^{(n_2 - n_1)}}{n_1!(n_2 - n_1)!}$

   (c) $\frac{e^{-\lambda t_2} {t_1}^{n_1} {t_2}^{n_2} (t_2 - t_1)^{(n_2 - n_1)}}{n_1!(n_2 - n_1)!}$ (d) none

3. The auto correlation function of t_1, t_2 of the Poisson process when $t_2 \geq t_1$ given by $R_{XX}(t_1, t_2) =$ _____

   (a) $\lambda^2 t_1 t_2 + \lambda t_1$ (b) $\lambda^2 t_2 + \lambda t_1$

   (c) $\lambda(t_1 + t_2)$ (d) none

4. The covariance function of the Poisson process for $t_2 \geq t_1$ is given by $C_{XX}(t_1, t_2) =$ _____

   (a) λt_2 (b) λt_1

   (c) $\lambda(t_1 + t_2)$ (d) none

5. The correlation coefficient function of the Poisson process for $t_2 > t_1$ is given by

   (a) $t_1 t_2$ (b) $\frac{t_2}{t_1}$

   (c) $\frac{t_1}{t_2}$ (d) none

6. The n-step transition probability $P_{ij}^{(n)}$ in state a_j at step n and given that it was in state a_i at step 0 is _____

   (a) $P\{X_o = a_i / X_n = a_j\}$ (b) $P\{X_n = a_j / X_o = a_i\}$

   (c) $P\{X_n = a_i / X_o = a_j\}$ (d) none

7. The state i of a Markov chain is called a return state if one of the following is true for $n \geq 1$

   (a) $P_{ij}^{(n)} = 1$ (b) $P_{ij}^{(n)} = 0$

   (c) $P_{ij}^{(n)} > 0$ (d) none

8. The period d_i of a return state i for $P_{ij}^{(m)} > 0$ is

   (a) GCD$\{m\}$ (b) LCM$\{m\}$

   (c) Product $\{m\}$ (d) none

9. The first return time probability of state i or the recurrence time probability F_n is given by

(a) $\sum_{n=1}^{\infty} f_{ii}^{(n)} = 1$ (b) $\pi_{n=1}^{\infty} f_{ii}^{(n)} = 1$

(c) $\frac{d}{dn} f_{ii}^{(n)} = 1$ (d) none

10. The mean recurrence time of the state i is given by $M_{ii} =$ _____

(a) $\sum_{n=1}^{\infty} n f_{ii}^{(n)}$ (b) $\sum_{n=1}^{\infty} f_{ii}^{(n)}$

(c) $\sum_{n=1}^{\infty} \frac{f_{ii}^{(n)}}{n} = 1$ (d) none

11. A state i is said to be persistent or recurrent if the return to state i is certain (i.e.,)

(a) $F_{ii} < 1$ (b) $F_{ii} > 1$

(c) $F_{ii} = 1$ (d) none

12. The covariance matrix of n random variables is given by $C_{ij} =$ _____ for $i \neq j$

(a) $E[(X_i - \bar{X}_j)(X_j - \bar{X}_i)]$ (b) $E[(X - \bar{X})^2]$

(c) $E[(X_i - \bar{X})(X_j - \bar{X}_j)]$ (d) none

13. If $R_X(t)$ is the envelope and θ_X is the phase of the process $X(t)$, then narrow band Gaussian process $X(t)$ is given by

(a) $\cos(\omega_o \pm \theta_X(t))$ (b) $R_X(t)\cos(\omega_o \pm \theta_X(t))$

(c) $R_X(t)\sin(\omega_o \pm \theta_X(t))$ (d) none

14. The band process is called band-limited process if the power spectrum of the band pass random process is _____ outside some frequency band of width ω that does not include $\omega = 0$

(a) zero (b) > 1

(c) <1 (d) none

ANSWERS

1. (c)	2. (b)	3. (a)	4. (b)	5. (c)	6. (b)	7. (c)	8. (a)
9. (a)	10. (a)	11. (c)	12. (c)	13. (b)	14. (a)		

This page is intentionally left blank.

Appendix A

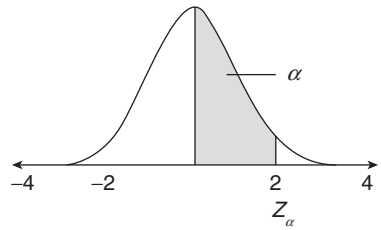


Table A Areas Under Standard Normal Distribution

	0	**0.01**	**0.02**	**0.03**	**0.04**	**0.05**	**0.06**	**0.07**	**0.08**	**0.09**
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990
3.1	0.4990	0.4991	0.4991	0.4991	0.4992	0.4992	0.4992	0.4992	0.4993	0.4993
3.2	0.4993	0.4993	0.4994	0.4994	0.4994	0.4994	0.4994	0.4995	0.4995	0.4995
3.3	0.4995	0.4995	0.4995	0.4996	0.4996	0.4996	0.4996	0.4996	0.4996	0.4997
3.4	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4998
3.5	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998
3.6	0.4998	0.4998	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999
3.7	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999
3.8	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999
3.9	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000
4.0	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000

Appendix B

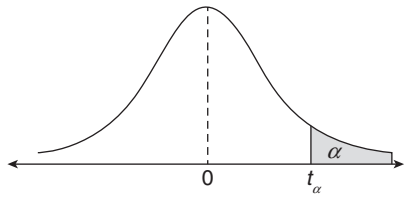


Table B Critical Values of the *t*-Distribution

	α						
υ	**0.40**	**0.30**	**0.20**	**0.15**	**0.10**	**0.05**	**0.025**
1	0.325	0.727	1.376	1.963	3.078	6.314	12.706
2	0.289	0.617	1.061	1.386	1.886	2.920	4.303
3	0.277	0.584	0.978	1.250	1.638	2.353	3.182
4	0.271	0.569	0.941	1.190	1.533	2.132	2.776
5	0.267	0.559	0.920	1.156	1.476	2.015	2.571
6	0.265	0.553	0.906	1.134	1.440	1.943	2.447
7	0.263	0.549	0.896	1.119	1.415	1.895	2.365
8	0.262	0.546	0.889	1.108	1.397	1.860	2.306
9	0.261	0.543	0.883	1.100	1.383	1.833	2.262
10	0.260	0.542	0.879	1.093	1.372	1.812	2.228
11	0.260	0.540	0.876	1.088	1.363	1.796	2.201
12	0.259	0.539	0.873	1.083	1.356	1.782	2.179
13	0.259	0.538	0.870	1.079	1.350	1.771	2.160
14	0.258	0.537	0.868	1.076	1.345	1.761	2.115
15	0.258	0.536	0.866	1.074	1.341	1.753	2.131
16	0.258	0.535	0.865	1.071	1.337	1.746	2.120
17	0.257	0.534	0.863	1.069	1.333	1.740	2.110
18	0.257	0.534	0.862	1.067	1.330	1.734	2.101
19	0.257	0.533	0.861	1.066	1.328	1.729	2.093
20	0.257	0.533	0.860	1.064	1.325	1.725	2.086
21	0.257	0.532	0.859	1.063	1.323	1.721	2.080
22	0.256	0.532	0.858	1.061	1.321	1.717	2.071
23	0.256	0.532	0.858	1.060	1.319	1.714	2.069
24	0.256	0.531	0.857	1.059	1.318	1.711	2.061
25	0.256	0.531	0.856	1.058	1.316	1.708	2.060
26	0.256	0.531	0.856	1.058	1.315	1.706	2.056
27	0.256	0.531	0.855	1.057	1.314	1.703	2.052
28	0.256	0.530	0.855	1.056	1.313	1.701	2.048
29	0.256	0.530	0.854	1.055	1.311	1.699	2.045
30	0.256	0.530	0.854	1.055	1.310	1.697	2.012
40	0.255	0.529	0.851	1.050	1.303	1.684	2.021
60	0.254	0.527	0.848	1.045	1.296	1.671	2.000
120	0.254	0.526	0.845	1.041	1.289	1.658	1.980
∞	0.253	0.524	0.842	1.036	1.282	1.645	1.960

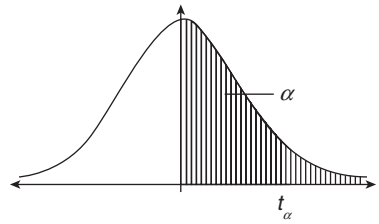


Table B Critical Values of the *t*-Distribution (*Continued*)

	α						
v	**0.02**	**0.015**	**0.01**	**0.0075**	**0.005**	**0.0025**	**0.0005**
1	15.895	21.205	31.821	42.434	63.657	127.322	636.578
2	4.849	5.643	6.965	8.073	9.925	14.089	31.598
3	3.482	3.896	4.541	5.047	5.841	7.453	12.924
4	2.999	3.298	3.747	4.088	4.604	5.598	8.610
5	2.757	3.003	3.365	3.634	4.032	4.773	6.869
6	2.612	2.829	3.143	3.372	3.707	4.317	5.959
7	2.517	2.715	2.998	3.203	3.499	4.029	5.408
8	2.449	2.634	2.896	3.085	3.355	3.833	5.041
9	2.398	2.574	2.821	2.998	3.250	3.690	4.781
10	2.359	2.527	2.764	2.932	3.169	3.581	4.587
11	2.328	2.491	2.718	2.879	3.106	3.497	4.437
12	2.303	2.461	2.681	2.836	3.055	3.428	4.318
13	2.282	2.436	2.650	2.801	3.012	3.372	4.221
14	2.264	2.415	2.624	2.771	2.977	3.326	4.140
15	2.249	2.397	2.602	2.746	2.947	3.286	4.073
16	2.235	2.382	2.583	2.724	2.921	3.252	4.015
17	2.224	2.368	2.567	2.706	2.898	3.222	3.965
18	2.214	2.356	2.552	2.689	2.878	3.197	3.922
19	2.205	2.346	2.539	2.674	2.861	3.174	3.883
20	2.197	2.336	2.528	2.661	2.845	3.153	3.849
21	2.189	2.328	2.518	2.649	2.831	3.135	3.819
22	2.183	2.320	2.508	2.639	2.819	3.119	3.792
23	2.177	2.313	2.500	2.629	2.807	3.104	3.768
24	2.172	2.307	2.492	2.620	2.797	3.091	3.745
25	2.167	2.301	2.485	2.612	2.787	3.078	3.725
26	2.162	2.296	2.479	2.605	2.779	3.067	3.707
27	2.158	2.291	2.473	2.598	2.771	3.057	3.690
28	2.154	2.286	2.467	2.592	2.763	3.047	3.674
29	2.150	2.282	2.462	2.586	2.756	3.038	3.659
30	2.147	2.278	2.457	2.581	2.750	3.030	3.646
40	2.125	2.250	2.423	2.542	2.704	2.971	3.551
60	2.099	2.223	2.390	2.504	2.660	2.915	3.460
120	2.076	2.196	2.358	2.468	2.617	2.860	3.373
∞	2.054	2.170	2.326	2.432	2.576	2.807	3.291

Appendix C

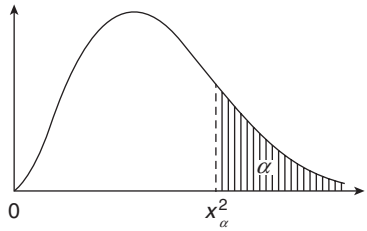


Table C Critical Values of the Chi-Square Distribution

	α									
v	**0.995**	**0.99**	**0.98**	**0.975**	**0.95**	**0.90**	**0.80**	**0.75**	**0.75**	**0.50**
1	0.0^4393	0.0^3157	0.0^3628	0.0^3982	0.00393	0.0158	0.0642	0.102	0.148	0.455
2	0.0100	0.0201	0.0404	0.0506	0.103	0.211	0.446	0.575	0.713	1.386
3	0.0717	0.115	0.185	0.216	0.352	0.584	1.005	1.213	1.424	2.366
4	0.207	0.297	0.429	0.484	0.711	1.064	1.649	1.923	2.195	3.357
5	0.412	0.554	0.752	0.831	1.145	1.610	2.343	2.675	3.000	4.351
6	0.676	0.872	1.134	1.237	1.635	2.204	3.070	3.455	3.828	5.348
7	0.989	1.239	1.564	1.690	2.167	2.833	3.822	4.255	4.671	6.346
8	1.344	1.646	2.032	2.180	2.733	3.490	4.594	5.071	5.527	7.344
9	1.735	2.088	2.532	2.700	3.325	4.168	5.380	5.899	6.393	8.343
10	2.156	2.558	3.059	3.247	3.940	4.865	6.179	6.737	7.267	9.342
11	2.603	3.053	3.609	3.816	4.575	5.578	6.989	7.584	8.148	10.341
12	3.074	3.571	4.178	4.404	5.226	6.304	7.807	8.438	9.034	11.340
13	3.565	4.107	4.765	5.009	5.892	7.042	8.634	9.299	9.926	12.340
14	4.075	4.660	5.368	5.629	6.571	7.790	9.467	10.165	10.821	13.339
15	4.601	5.229	5.985	6.262	7.261	8.547	10.307	11.036	11.721	14.339
16	5.142	5.812	6.614	6.908	7.962	9.312	11.152	11.912	12.624	15.338
17	5.697	6.408	7.255	7.564	8.672	10.085	12.002	12.792	13.531	16.338
18	6.265	7.015	7.906	8.231	9.390	10.865	12.857	13.675	14.440	17.338
19	6.844	7.633	8.567	8.907	10.117	11.651	13.716	14.562	15.352	18.338
20	7.434	8.260	9.237	9.591	10.851	12.443	14.578	15.452	16.266	19.337
21	8.034	8.897	9.915	10.283	11.591	13.240	15.445	16.344	17.182	20.337
22	8.643	9.542	10.600	10.982	12.338	14.041	16.314	17.240	18.101	21.337
23	9.260	10.196	11.293	11.688	13.091	14.848	17.187	18.137	19.021	22.337
24	9.886	10.856	11.992	12.401	13.848	15.659	18.062	19.037	19.943	23.337
25	10.520	11.524	12.697	13.120	14.611	16.473	18.940	19.939	20.867	24.337
26	11.160	12.198	13.409	13.844	15.379	17.292	19.820	20.843	21.792	25.336
27	11.808	12.879	14.125	14.573	16.151	18.114	20.703	21.749	22.719	26.336
28	12.461	13.565	14.847	15.308	16.928	18.939	21.588	22.657	23.647	27.336
29	13.121	14.256	15.574	16.047	17.708	19.768	22.475	23.567	24.577	28.330
30	13.787	14.953	16.306	16.791	18.493	20.599	23.364	24.478	25.508	29.336
40										
50										
60										

Table C Critical Values of the Chi-Square Distribution (*Continued*)

	α									
υ	**0.30**	**0.25**	**0.20**	**0.10**	**0.05**	**0.025**	**0.02**	**0.01**	**0.005**	**0.001**
1	1.074	1.323	1.642	2.706	3.841	5.024	5.412	6.635	7.879	10.827
2	2.408	2.773	3.219	4.605	5.991	7.378	7.824	9.210	10.597	13.815
3	3.665	4.108	4.642	6.251	7.815	9.348	9.837	11.345	12.838	16.268
4	4.878	5.385	5.989	7.779	9.488	11.143	11.668	13.277	14.860	18.465
5	6.064	6.626	7.289	9.236	11.070	12.832	13.388	15.086	16.750	20.517
6	7.231	7.841	8.558	10.645	12.592	14.449	15.033	16.812	18.548	22.457
7	8.383	9.037	9.803	12.017	14.067	16.013	16.622	18.475	20.278	24.322
8	9.524	10.219	11.030	13.362	15.507	17.535	18.168	20.090	21.955	26.125
9	10.656	11.389	12.242	14.684	16.919	19.023	19.679	21.666	23.589	27.877
10	11.781	12.549	13.442	15.987	18.307	20.483	21.161	23.209	25.188	29.588
11	12.899	13.701	14.631	17.275	19.675	21.920	22.618	24.725	26.757	31.264
12	14.011	14.845	15.812	18.549	21.026	23.337	24.054	26.217	28.300	32.909
13	15.119	15.984	16.985	19.812	22.362	24.736	25.472	27.688	29.819	34.528
14	16.222	17.117	18.151	21.064	23.685	26.119	26.873	29.141	31.319	36.123
15	17.322	18.245	19.311	22.307	24.996	27.488	28.259	30.578	32.801	37.697
16	18.418	19.369	20.465	23.542	26.296	28.845	29.633	32.000	34.267	39.252
17	19.511	20.489	21.615	24.769	27.587	30.191	30.995	33.409	35.718	40.790
18	20.601	21.605	22.760	25.989	28.869	31.526	32.346	34.805	37.156	42.312
19	21.689	22.718	23.900	27.204	30.144	32.852	33.687	36.191	38.582	43.820
20	22.775	23.828	25.038	28.412	31.410	34.170	35.020	37.566	39.997	45.315
21	23.858	24.935	26.171	29.615	32.671	35.479	36.343	38.932	41.401	46.797
22	24.939	26.039	27.301	30.813	33.924	36.781	37.659	40.289	42.796	48.268
23	26.018	27.141	28.429	32.007	35.172	38.076	38.968	41.638	44.181	49.728
24	27.096	28.241	29.553	33.196	36.415	39.364	40.270	42.980	45.558	51.179
25	28.172	29.339	30.675	34.382	37.652	40.646	41.566	44.314	46.928	52.620
26	29.246	30.434	31.795	35.563	38.885	41.923	42.856	45.642	48.290	54.052
27	30.319	31.528	32.912	36.741	40.113	43.194	44.140	46.963	49.645	55.476
28	31.391	32.620	34.027	37.916	41.337	44.461	45.419	48.278	50.993	56.893
29	32.461	33.711	35.139	39.087	42.557	45.722	46.693	49.588	52.336	58.302
30	33.530	34.800	36.250	40.256	43.773	46.979	47.962	50.892	53.672	59.703

Appendix D

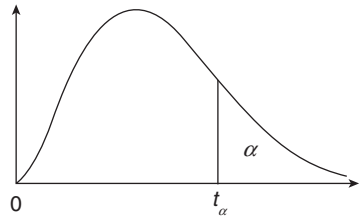


Table D* Critical Values of the *F*-Distribution

$F_{0.05}(v_1, v_2)$

	v_1								
v_2	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
1	161.4	199.5	215.71	224.16	230.16	234.0	236.77	238.88	240.54
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45.	2.39
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04
120	3.92	3.07	2.68	2.45	2.29	2.17	2.09	2.02	1.96
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88

Table D Critical Values of the *F*-Distribution (*Continued*)

$F_{0.05}(v_1, v_2)$

	v_1									
v_2	**10**	**12**	**15**	**20**	**24**	**30**	**40**	**60**	**120**	**∞**
1	241.9	243.91	245.95	248.01	249.05	250.1	251.1	252.2	253.25	254.31
2	19.40	19.41	19.43	19.45	19.45	19.46	19.47	19.48	19.49	19.50
3	8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53
4	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
5	4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.36
6	4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67
7	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23
8	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93
9	3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71
10	2.98	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.54
11	2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.40
12	2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.30
13	2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21
14	2.60	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13
15	2.54	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.07
16	2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.01
17	2.45	2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01	1.96
18	2.41	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	1.92
19	2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.88
20	2.35	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	1.84
21	2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87	1.81
22	2.30	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	1.78
23	2.27	2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.81	1.76
24	2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79	1.73
25	2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.71
26	2.22	2.15	2.07	1.99	1.95	1.90	1.85	1.80	1.75	1.69
27	2.20	2.13	2.06	1.97	1.93	1.88	1.84	1.79	1.73	1.67
28	2.19	2.12	2.04	1.96	1.91	1.87	1.82	1.77	1.71	1.65
29	2.18	2.10	2.03	1.94	1.90	1.85	1.81	1.75	1.70	1.64
30	2.16	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	1.62
40	2.08	2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.58	1.51
60	1.99	1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47	1.39
120	1.91	1.83	1.75	1.66	1.61	1.55	1.50	1.43	1.35	1.25
∞	1.83	1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22	1.00

(*Continued*)

Table D Critical Values of the *F*-Distribution (*Continued*)

	$F_{0.01}(v_1, v_2)$								
	v_1								
v_2	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
1	4052.18	4999.5	5403.35	5624.58	5763.65	5858.99	5928.36	5981.07	6022.47
2	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39
3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35
4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66
5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16
6	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98
7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72
8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91
9	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35
10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78
17	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26
25	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22
26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18
27	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15
28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12
29	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72
120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56
∞	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41

Table D Critical values of the *F*-Distribution (*Continued*)

	$F_{0.01}(v_1, v_2)$									
	v_1									
v_2	**10**	**12**	**15**	**20**	**24**	**30**	**40**	**60**	**120**	**∞**
1	6055.85	6106.32	6157.28	6208.73	6234.63	6260.65	6286.78	6313.03	6339.39	6365.86
2	99.40	99.42	99.43	99.45	99.46	99.47	99.47	99.48	99.49	99.50
3	27.23	27.05	26.87	26.69	26.60	26.50	26.41	26.32	26.22	26.13
4	14.55	14.37	14.20	14.02	13.93	13.84	13.75	13.65	13.56	13.46
5	10.05	9.89	9.72	9.55	9.47	9.38	9.29	9.20	9.11	9.02
6	7.87	7.72	7.56	7.40	7.31	7.23	7.14	7.06	6.97	6.88
7	6.62	6.47	6.31	6.16	6.07	5.99	5.91	5.82	5.74	5.65
8	5.81	5.67	5.52	5.36	5.28	5.20	5.12	5.03	4.95	4.86
9	5.26	5.11	4.96	4.81	4.73	4.65	4.57	4.48	4.40	4.31
10	4.85	4.71	4.56	4.41	4.33	4.25	4.17	4.08	4.00	3.91
11	4.54	4.40	4.25	4.10	4.02	3.94	3.86	3.78	3.69	3.60
12	4.30	4.16	4.01	3.86	3.78	3.70	3.62	3.54	3.45	3.36
13	4.10	3.96	3.82	3.66	3.59	3.51	3.43	3.34	3.25	3.17
14	3.94	3.80	3.66	3.51	3.43	3.35	3.27	3.18	3.09	3.00
15	3.80	3.67	3.52	3.37	3.29	3.21	3.13	3.05	2.96	2.87
16	3.69	3.55	3.41	3.26	3.18	3.10	3.02	2.93	2.84	2.75
17	3.59	3.46	3.31	3.16	3.08	3.00	2.92	2.83	2.75	2.65
18	3.51	3.37	3.23	3.08	3.00	2.92	2.84	2.75	2.66	2.57
19	3.43	3.30	3.15	3.00	2.92	2.84	2.76	2.67	2.58	2.49
20	3.37	3.23	3.09	2.94	2.86	2.78	2.69	2.61	2.52	2.42
21	3.31	3.17	3.03	2.88	2.80	2.72	2.64	2.55	2.46	2.36
22	3.26	3.12	2.98	2.83	2.75	2.67	2.58	2.50	2.40	2.31
23	3.21	3.07	2.93	2.78	2.70	2.62	2.54	2.45	2.35	2.26
24	3.17	3.03	2.89	2.74	2.66	2.58	2.49	2.40	2.31	2.21
25	3.13	2.99	2.85	2.70	2.62	2.54	2.45	2.36	2.27	2.17
26	3.09	2.96	2.81	2.66	2.58	2.50	2.42	2.33	2.23	2.13
27	3.06	2.93	2.78	2.63	2.55	2.47	2.38	2.29	2.20	2.10
28	3.03	2.90	2.75	2.60	2.52	2.44	2.35	2.26	2.17	2.06
29	3.00	2.87	2.73	2.57	2.49	2.41	2.33	2.23	2.14	2.03
30	2.98	2.84	2.70	2.55	2.47	2.39	2.30	2.21	2.11	2.01
40	2.80	2.66	2.52	2.37	2.29	2.20	2.11	2.02	1.92	1.80
60	2.63	2.50	2.35	2.20	2.12	2.03	1.94	1.84	1.73	1.60
120	2.47	2.34	2.19	2.03	1.95	1.86	1.76	1.66	1.53	1.38
∞	2.32	2.18	2.04	1.88	1.79	1.70	1.59	1.47	1.32	1.00

*Reproduced from Table 18 of *Biometrika Tables for Statisticians*, Vol. I, by permission of E. S. Pearson and the Biometrika Trustee.

This page is intentionally left blank.

Index

This page is intentionally left blank.