PROBABILITY & STATISTICS

M.Sc (COMPUTER SCIENCE)
SEMESTER-I, PAPER-IV

Lesson Writers:

Dr. A.J.V. Radhika
Asst.Professor,
Dept. of Basic Science & Humanities,
Acharya Nagarjuna University,
Nagarjunanagar – 522 510

Dr. M. Syam Sundar
Asst. Professor
Dept. of Nano Technology
Acharya Nagarjuna University
Nagarjunanagar – 522 510

Dr. T.V. Pradeep Kumar
Asst. Professor
Dept. of Basic Science & Humanities
Acharya Nagarjuna University
Nagarjunanagar – 522 510

Dr. B. Sri Ram
Faculty
Dept. of Basic Science & Humanities
Acharya Nagarjuna University
Nagarjunanagar – 522 510

Editor

Dr. V. Amarendra Babu
Asst. Professor,
Dept. Of Mathematics,
Acharya Nagarjuna University
Nagarjunanagar – 522 510.

Director, I/c.
Prof. V. Venkateswarlu
M.A., M.P.S., M.S.W ., M.Phil., Ph.D.
Professor
Centre for Distance Education
Acharya Nagarjuna University
Nagarjuna Nagar 522 510

Ph: 0863-2346222, 2346208
0863- 2346259 (Study Material)
Website www.anucde.info
E-mail: anucdedirector@gmail.com

M.Sc Computer Science

First Edition : 2025

No. of Copies :

This book is exclusively prepared for the use of students of M.Sc (Computer Science), Centre for Distance Education, Acharya Nagarjuna University and this book is meant for limited circulation only.

Published by:

Director **I/c**
Prof. V. Venkateswarlu,
M.A., M.P.S., M.S.W . M.Phil., Ph.D.
Centre for Distance Education,
Acharya Nagarjuna University

Printed at:

FOREWORD

Since its establishment in 1976, Acharya Nagarjuna University has been forging ahead in the path of progress and dynamism, offering a variety of courses and research contributions. I am extremely happy that by gaining 'A+' grade from the NAAC in the year 2024, Acharya Nagarjuna University is offering educational opportunities at the UG, PG levels apart from research degrees to students from over 221 affiliated colleges spread over the two districts of Guntur and Prakasam.

The University has also started the Centre for Distance Education in 2003-04 with the aim of taking higher education to the door step of all the sectors of the society. The centre will be a great help to those who cannot join in colleges, those who cannot afford the exorbitant fees as regular students, and even to housewives desirous of pursuing higher studies. Acharya Nagarjuna University has started offering B.Sc., B.A., B.B.A., and B.Com courses at the Degree level and M.A., M.Com., M.Sc., M.B.A., and L.L.M., courses at the PG level from the academic year 2003-2004 onwards.

To facilitate easier understanding by students studying through the distance mode, these self-instruction materials have been prepared by eminent and experienced teachers. The lessons have been drafted with great care and expertise in the stipulated time by these teachers. Constructive ideas and scholarly suggestions are welcome from students and teachers involved respectively. Such ideas will be incorporated for the greater efficacy of this distance mode of education. For clarification of doubts and feedback, weekly classes and contact classes will be arranged at the UG and PG levels respectively.

It is my aim that students getting higher education through the Centre for Distance Education should improve their qualification, have better employment opportunities and in turn be part of country's progress. It is my fond desire that in the years to come, the Centre for Distance Education will go from strength to strength in the form of new courses and by catering to larger number of people. My congratulations to all the Directors, Academic Coordinators, Editors and Lesson-writers of the Centre who have helped in these endeavors.

Prof. K. Gangadhara Rao
M.Tech., Ph.D.,
Vice-Chancellor I/c
Acharya Nagarjuna University

M.Sc. Computer Science
Semester-I, Paper-IV
Probability & Statistics

Syllabus

UNIT 1:

Some probability laws: Axioms of Probability, Conditional Probability, Independence of the Multiplication Rule, Bayes' theorem

Discrete Distributions: Random Variables, Discrete Probability Densities, Expectation and distribution parameters, Binomial distribution, Poisson distribution, simulating a Discrete distribution

UNIT II:

Continuous distributions: continuous Densities, Expectation and distribution parameters, exponential distribution, Normal distribution, Weibull distribution and Reliability. Estimation: Point estimation, interval estimation and central limit theorem.

UNIT III:

Inferences on the mean and the Variance of a distribution: Hypothesis Testing, significance testing, Hypothesis and significance test on the mean, Hypothesis tests on the Variance Inferences on proportions: estimating proportions, testing hypothesis on a proportion, Comparing two proportions: estimation, comparing two proportions: hypothesis testing.

UNIT IV:

Comparing two means and two variances: point estimation: independent samples, Comparing variances: the F-distribution, Comparing means: variances equal, Analysis of Variance: One-way classification fixed effects model, comparing variances, pair wise comparisons, randomized complete block design

UNIT V:

Simple linear regression and correlation: model and parameter estimation, inferences about slope, inferences about intercept, Co-efficient of determination Multiple linear regression models: least square procedures for model fitting, a matrix approach to least squares, interval estimation.

Prescribed book:

J Susan Milton and Jesse C. Arnold: "Introduction to Probability and Statistics", Fourth edition, TMH, (2007).

Reference book:

William Mendenhall, Robert J Beaver, Barbara M Beaver: Introduction to Probability and Statistics, Twelth edition, Thomson.

(104CP24)

M.SC DEGREE EXAMINATION, Model QP

Computer Science – First Semester

Probability & Statistics

Time: 3hours Max. Marks: 70

Answer ONE Question from each unit 5 x 14-70 M

UNIT – I

1. a) Explain conditional probability.
   b) Justify the need for normal distribution.

(or)

2. Write a routine to plot a histogram that compares binomial and normal distribution

UNIT-II

3. Fit a normal distribution to the following data and also test the adequacy of the model.

X:	0	1	2	3	4	5
Y:	3	9	12	27	4	1

(or)

4. a) Fit a straight line y=a+bx to the following data:

X:	12	17	19	25	32	38
Y:	65	78	82	92	90	97	100

Also estimate y when x=35.

UNIT-III

5. a) Derive the test statistic on F-test.
   b) Explain the statistical analysis of one way classification.

(OR)

6. Two random samples drawn from two normal populations are

Sample 1	20	16	26	27	23	22	18	24	25	19
Sample 2	17	33	42	35	32	34	38	28	41	43

Obtain the estimates of the variances of the population. Test whether the populations have same variances.

UNIT-IV

7. It is claimed that a random sample of 49 tyres has a mean life of 15200 km. The sample was drawn from a population whose mean in 151150 km and a standard deviation of 1200 km. Test the significance at 0.05 level also find 95% confidence limits.

(OR)

8. a)Why ANOVA is used for comparisons of multiple means instead of multiple test?
   b) Discuss the ANOVA Model for RBD.

UNIT-V

9. a) Explain a matrix approach to least squares in multiple linear regression models.
   b) Explain i)Coefficient of determination ii) Interval estimation.

(OR)

10. a)Explain the inferences about Slope and intercept.
    b) The two lines of regression are 8x-10y+66=0, 40x-18y-214=0. The variance of x is 9. Find (i) the mean values of x and y (ii) Correlation coefficient between x and y.

CONTENTS

TITLE	PAGE NO
1. Probability	1.1- 1.16
2. Conditional Probability	2.1- 2.13
3. Random Variables	3.1- 3.15
4. Discrete Distribution	4.1- 4.22
5. Continuous Probability Distribution	5.1- 5.13
6. Normal Distribution	6.1- 6.20
7. Weibull Distribution	7.1- 7.11
8. Estimation	8.1- 8.18
9. Inference on the mean	9.1- 9.16
10. Inference on the Variance	10.-10.14
11. Inference on one Proportion	11.1-11.12
12. Inference on Two Proportions	12.1-12.15
13. Comparing Two Means	13.1- 13.19
14. Comparing Two Variances	14.1- 14.21
15. Analysis of Variance	15.1- 15.15
16. Randomised Complete Block Design	16.1 -16.20
17. Simple Linear Regression	17.1 -17.29
18. Correlation	18.1-18.19
19. Matrix Notation	19.1-19.22
20. Interval Estimation	20.1-20.07

LESSON- 1

PROBABILITY

OBJECTIVES:

After going through this lesson, you will be able to

- Understanding the Fundamental Concepts of Probability
- Explore Axioms of probability
- Analyze Characteristic Equations
- Examine elementary theorems on probability

STRUCTURE OF THE LESSION:

1.1 Introduction

1.2 Basic Terminology Definition of probability.

1.3 Definition of probability.

1.4 worked out examples

1.5 Axioms of probability.

1.6 Some elementary theorems on probability

1.7 Worked out problems

1.8 Summary

1.9 **Technical Terms**

1.10 Self Assessment Questions

1.11 **Further Readings**

1.1 INTRODUCTION

Probability is the branch of Mathematics concerning events and numerical description of how likely they are to occur. There are many real life situations in which we may have to predict the outcome of an event. We may be sure (or) not sure of the results of an event. In such cases, we say that there is a probability of this event to occur (or) not occur.

The words like “possibly”, “high chance”, “likely” and "odds” are expressions indicating a degree of uncertainty about the happening of an event. i.e A numerical measure of uncertainty is provided by a branch of Mathematics called the "Theory of probability”. Broadly, there are three possible states of expectation – ‘certainty’, impossibility and Uncertainty.

The probability theory describes certainty by 1, impossibility by 0 and the various grades of uncertainties by coefficients ranging between 0 and 1.

1.2. BASIC TERMINOLOGY

- **Random Experiment:** An experiment (or) trial is any procedure that can be infinitely repeated and has a well-defined set of possible outcomes, known as the sample space. The sample space denoted by S.

An experiment is said to be random if it has more than one possible outcomes. Examples of random experiments are:

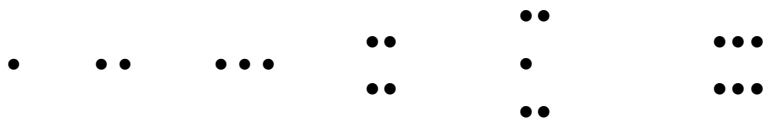
Tossing a coin, throwing a die, selecting a card from a pack of playing cards etc. In all these cases there are a number of possible results which can occur but there is an uncertainty as to which one of them will actually occur.

Note:

(i)Tossing a coin, the possible outcomes head (H) tail (T)

(ii)Throwing a die, the possible outcomes. are 1, 2, 3, 4, 5, 6.

i.e, A die is a small cube, on its six faces dots are marked as



Plural of die is dice, the outcome of throwing a die is the number of dots on its uppermost face.

(iii) A pack of cards consists of four suits called Spades, Hearts, Diamonds and clubs. Each suit consists of 13 cards, of which 9 cards are numbered from 2 to 10, an ace, king, and a queen and a jack (or knave).
Spades and Clubs are black-faced cards while Hearts and Diamonds are red-faced cards.

- **Outcome:** The result of a random experiment is called outcome.
- **Trial:** Any particular performance of a random experiment is called a trial.
- **Event:** Outcome (or) combination of outcomes are termed as events.

For example:

1. In Tossing a coin is trial, turned up Head (or) Tail is an outcome. There are two Possible cases either Head (H) (or) Tail (T) Sample Space(S) is { H,T}

2. In throwing a die, there are six possibilities 1,2 3, 4, 5,6

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

3. In tossing two coins (or) one coin two times the possible outcomes are { HH, HT, TH,TT }

4. In tossing 3 coins (or) one coin 3 times, Sample Space

S={HHH, HHT, HTH,THH, HTT THT, TTH,TTT }

Note: -

(i) If a coin tossed n times (or) n coins tossed at a time, the sample space contains 2^n elements.

(ii) If a die thrown for n times (or) n dice thrown, the sample space consists of 6^n elements.

5. In throwing a die two times (or) two dice at a time, the sample space consists of 36 elements.

$$S=\begin{pmatrix} (1,1),(1,2),\ (1,\ 3),(1,4),(1,5),(1,6) \\ (2,1),(2,2),(2,3),(2,4),(2,5),(2,6) \\ (3,1),(3,2)\ \ (3,3),(3,4),(3,5),\ (3,6) \\ (4,1),(4,2),(4,3),(4,4),(4,5),(4,6) \\ (5,1),(5,2),(5,3),(5,4),(5,5),(5,6) \\ (6,1),(6,2),(6,3),(6,4),(6,5),(6,6) \end{pmatrix}$$

- **Favourable cases:** The number of cases favourable to an event in a trial is the number of outcomes which entail the happening of an event.

Ex:

(1) In drawing a card from a Pack of 52 cards, the favourable cases in drawing a Queen is 4.

2) In drawing a Red Card is 26

3) In throwing two dice, the number of cases favourable in getting a sum 7 is 6.

i.e , (1,6) (2,5) (3,4) (4,3) (5,2) (6,1).

- **Exhaustive Events (or) Cases:** The Total number of possible outcomes of a random experiment is known as the exhaustive events (or) cases.

For example

(i) In tossing a coin, there are two exhaustive cases; head and tail

(ii)In throwing of a die, there are 6 exhaustive Cases

(iii)In drawing two cards from a pack of cards

The exhaustive number of cases is $^{52}C_2$ Since 2 cards can be drawn out of 52 cards in $^{52}C_2$ ways.

- **Mutually exclusive events:** Events are said to be mutually exclusive if the happening of anyone of them precludes the happening of all the others. i.e, if no two (or) more of them can happen simultaneously in the same Trial.

Ex:

(i)In tossing a coin, the events head and tail are mutually exclusive.

(ii)In throwing a die, all 6 possible cases are mutually exclusive.

Note: If two events are mutually exclusive, then the sets are disjoint, i.e if A and B

are mutually exclusive $A \cap B = \phi$;

In the above example, $[H] \cap [T] = \phi$

Ex:(i) A = [HH] , B= [HT,TH], C= [TT]

Here $A \cap B = \phi$, B$\cap$ C $= \phi$, $C \cap A = \phi$.

- **Equally likely events:** Outcomes of trial are said to be Equally likely events if one cannot be expected in preference to the others.

Ex:

(1) In throwing a dice all possible cases are equally likely events

(2). In tossing a coin, two possible cases are equally likely!

1.2 Probability:

Definition: If an experiment is performed n is the number of exhaustive cases and m is the number of favourable cases of an event A. Then probability of an event A is defined by

$$\frac{m}{n} = \frac{\text{number of favourable cases}}{\text{number of exhaustive cases}}$$

$$= \frac{n(\text{A})}{n(S)}$$

Where n(A)=Number of elements belonging to A

N(S)=Number of elements belonging to S(Sample Space).

1.3 WORKED OUT EXAMPLES

Problems:

(1) Find the probability of getting one head in tossing two coins.

Sol: Let A be the event of getting one head.

A={ HT,TH }

S= { HH, HT, TH,TT }

number of elements in A = 2

number of elements in S =4

$$\therefore P(A) = \frac{2}{4} = \frac{1}{2}$$

(2) Find the probability of getting one red King if we select a card from pack of 52 cards.

Sol: There are 2 red kings

Number of possible cases = 2

Number of exhaustive cases = 52

$$\text{Probability } P(A) = \frac{2}{52} = \frac{1}{26}$$

(3) Find the probability of getting a sum 9 if two dice are thrown.

Sol:

$$\text{The sample space S is} = \begin{pmatrix} (1,1),(1,2),\ (1,\ 3),(1,4),(1,5),(1,6) \\ (2,1),(2,2),(2,3),(2,4),(2,5),(2,6) \\ (3,1),(3,2)\ (3,3),(3,4),(3,5),\ (3,6) \\ (4,1),(4,2),(4,3),(4,4),(4,5),(4,6) \\ (5,1),(5,2),(5,3),(5,4),(5,5),(5,6) \\ (6,1),(6,2),(6,3),(6,4),(6,5),(6,6) \end{pmatrix}$$

Let A be the event of getting a sum of 9={ (3,6)(4,5)(5,4)(6,3) }

$$P(A) = \frac{4}{36} = \frac{1}{9}.$$

(4)If three coins are tossed. Find the probability of getting (i)Three heads (ii)Two heads (iii)no heads.

Sol:

(i) The sample space S = { HHH, HHT, THH, HTH, HIT,THT, TTH, TTT}

Let the event A be of getting three heads A = {HHH}

P(A) = 1/8

(ii) Let the event B be of getting two heads B ={ HHT, THH, HTH }

P(B) = 3/8

(iii)no heads

Let the event c be getting no heads

C={ TTT }

P(C)=1/8

(5) Find the probability of getting 2 diamonds, if we draw 2 cards at random from a pack of cards.

Sol: There are 52 cards total.

There are $52C_2$ ways to choose 2 cards.

Number of Exhaustive cases = $52C_2$ = 1326

There are $13C_2$ ways to chose 2 diamonds, since there are 13 diamonds.

No. of possible cases = $13C_2$ = 78

$$P(A) = \frac{78}{1376} = \frac{1}{17}$$

(6) Three cards are drawn from a pack of 52 cards. Find the probability that

(i) 3 are spades

(ii)2 spades ,one diamond.

(iii)1spade, 1 diamond, 1 heart

Sol:

(i) There are $52C_3$ ways to draw 3 cards from 52 Cards.

Number of Exhaustive cases= $52C_3$ =22100 Number of possible cases = $13C_3$ = 286

$$P(A) = \frac{286}{22100} = 143/11050 = 11/850$$

(ii)Number of exhaustive cases=22100

Number of possible cases= $13C_2 \times 13C_1$

$$P(A) = \frac{78 \times 13}{22100} = \frac{39}{850}$$

(iii) Number of Exhaustive cases =22100

Number of Exhaustive cases = $13_{C_1} \times 13_{C_1} \times 13_{C_1} = 2197$.

P(A) = 2197/22100 = 169/1700

(7)Three light bulbs are chosen at random from 12 bulbs of which 5 are defective. Find the probability that (i)All are defective (ii) one is defective (iii) two are defective

Sol: Number of Exhaustive cases = $12_{C_3} = 220$

(i) Number of possible cases = $5_{C_3} = 10$

Required probability = 10/220 = 1/22

(ii)Number of possible cases = $5_{C_1} \times 7_{C_2}$ (Since there are 7 Non defective bulbs)=105

Required probability =105/220 = 21/44

(Since there are 7 Non defective bulbs) = 105/220 = 21/44

(iii) Number of possible cases = $5_{C_2} \times 7_{C_1}$ =70

Required probability = 70/220 = 7/22

(8) what is the probability of drawing an ace from a well shuffled pack of 252 playing cards?

Sol: The number of exhaustive cards= 52_{C_1} =52

The number of possible cases= 4_{C_1} =4(Since there are 4 aces)

Probability = 4/52 = 1/13

(9) A bag contains 5 red balls, 8 blue balls and l1 white balls. Three balls are drawn together from the box. Find the probability that

(i) one is red, one is blue and one is white

(ii) Two white and one red

(iii) Three white

Sol: (i)Number of Exhaustive cases $24_{C_3} = 2024$

Number of possible cases = $5_{C_1} \times 8_{C_1} \times 11_{C_1}$ =440

Required probability =440/2024 = 55/253

(ii)Number of possible cases $= 11_{C_2} \times 5_{C_1} = 275$

Required probability = 275/2024 = 25/184

(iii)Number of possible cases=11_{C_3}

Required probability -= 165/2024 =15/184.

1.4 THE AXIOMS OF PROBABILITY

The Axioms of probability are

1.$0 \leq P(A) \leq 1$ for each event $A \subseteq S$

2. P(S)=1

3. If A and B are any two mutually exclusive events then

$P(A \cup B) = P(\mathrm{A}) + \mathrm{P(B)}$

Ex:

(1)Tossing a coin, Sample space S={ H,T }

P(H)=1/2 , P(T)=1/2 < 1

P(S)=P(H)+P(T)=1/2+1/2=1

{H},{T} are mutually exclusive.

$P[(\mathrm{H}) \cup (T)] = P\{H, T\} = P(S) = 1$

$P[(\mathrm{H}) \cup (T)] = P(\mathrm{A}) + \mathrm{P}(T) = 1/2 + 1/2 = 1$

(2) In throwing a die S={1,2,3,4,5,6}

P(1)=1/6,P(2)=1/6, P(3)=1/6,P(4)=1/6,P(5)=1/6,P(6)=1/6

P(1)+ P(2)+P(3)+ P(4)+ P(5)+P(6)=1

P(S)=1

Here all are mutually exclusive cases.

(3) In tossing two coins

Sample Space S={ HH,HT,TH,TT }

P(HH)=1/4

P(HT,TH)=1/2

P(TT)=1,4

P(HH)+P(HT,TH)+P(TT)=P(S)=1

These are mutually exclusive cases.

1.5. SOME THEOREMS ON PROBABILITY

In this section, we shall prove a few simple theorems which help us to evaluate the probabilities of some complicated events in a rather simple way. In proving these theorems we shall follow the axiomatic approach based on the three axioms.

Theorem1: Probability of the impossible event is zero, i.e, $P(\phi) = 0$

Proof : Impossible event contains no sample point and hence certain event S and the impossible event ϕ are mutually exclusive.

$S \cup \phi$ =S $\Longrightarrow$P($S \cup \phi$)=P(S)

Hence by using Axiom 2 of probability, i.e Axiom of Additivity, we get

P(S)+ $P(\phi)$ =P(S)

$\Longrightarrow$ $P(\phi) = 0$.

Theorem 2: Probability of the complementary event $\overline{A}$ of A is given by P($\overline{A}$)=1-P(A)

Proof: A and $\overline{A}$ are mutually disjoint events, so that $A \cup \overline{A} = S \Longrightarrow \mathrm{P}(A \cup \overline{A}) = P(S)$

Hence by using Axioms 2 and 3 of probability, P(A)+P($\overline{A}$)=P(S)=1

P($\overline{A}$)=1-P(A)

Theorem 3: For any two events A and B, We have

(i)**(** $P(\overline{A} \cap B) = P(B) - P(A \cap B)$

(ii) $P(\mathrm{A} \cap \overline{B}) = P(\mathrm{A}) - P(A \cap B)$

Proof: From the venn diagram

We get B= $B = (A \cap B) \cup (\overline{A} \cap B)$

Where $\overline{A} \cap B$ and $A \cap B$ are disjoint events. Hence by Axiom(3)

$P(B) = P(A \cap B) + \mathrm{P}(\overline{A} \cap B)$

$\mathrm{P}(\overline{A} \cap B) = P(B) + \mathrm{P}(\mathrm{A} \cap B)$

Similarly $\mathrm{P}(\mathrm{A} \cap \overline{B}) = P(\mathrm{A}) - \mathrm{P}(\mathrm{A} \cap B)$.

Addition theorem of Probability

Theorem: If A and B are any two events(Subsets of sample space S) and are not disjoint, then $P(A\cup B) = P(A) + P(B) - P(A\cap B)$

Proof: From the venn diagram, we have $A\cup B = A\cup(\overline{A}\cap B)$

Where A and $\overline{A}\cap B$ are mutually disjoint.

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

$= P(A) + P(\overline{A}\cap B)$ (By Axiom 3)

$= P(A) + P(B) - P(A\cap B)$ (From theorem (3))

Cor.1: If the events A and B are mutually disjoint

Then $A\cap B = \phi \Longrightarrow P(A\cap B) = P(\phi) = 0$

$P(A\cup B) = P(A) + P(B)$, which is Axiom (3) of probability

Cor.2: For three non-mutually exclusive events A, Band C , we have

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

Example

(1): The probability that a student passes a physics test is $\frac{2}{3}$ and the probability that he passes both a Physics test and an English test is $\frac{14}{45}$. The probability that he passes at least one test is $\frac{4}{5}$. What is the probability that he passes the English test?

Sol: A: The student passes a Physics test ,

B: The student passes a English test.

P(A)= 2/3 $P(A\cap B)$ = 14/45 : P(AUB) = 4/5 and P(B) = ?

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

4/5 = 2/3 + P(B) - 14/45

P(B) = 4/5 + 14/45 - 2/3 = (36 + 14 - 30)/45 = 4/9

Example (2):

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

Solution:

A: the card drawn is a king

B: the card drawn is a heart

C: the card drawn is a red card.

Then A, B and C are not mutually exclusive.

$A \cap B$: The card drawn is the king of hearts

$\Rightarrow n(A \cap B) = 1$

$B \cap C = B$; The card drawn is a heart $(B \subset C)$

$\Rightarrow n(B \cap C) = 13$

$C \cap A$: The card drawn is a red king of hearts

$\Rightarrow n(C \cap A) = 2$

$\Rightarrow n(A \cap B \cap C) = 1$

$\therefore P(A) = 4/52,\ P(B) = 13/52,\ P(C) = 26/52$, $P(A \cap B) = 1/52$, $P(B \cap C) = 13/52$,

$P(C \cap A) = 2/52$, $P(A \cap B \cap C) = 1/52$

The required probability of getting a king (or) heart (or) a red card is given by

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

$= 4/52 + 13/52 + 26/52 - 1/52 - 13/52 - 2/52 + 1/52 = 28/52 = 7/13.$

1.6 WORKED OUT PROBLEMS:

(1) What is the probability of a Leap year to have 52 Mondays and 53 Sundays?

Sol: A Leap year has 366 days i.e 52 weeks 2 days. These 2 days Cam be any one of the following 7 ways

1 Monday Tuesday

2 Tuesday Wednesday

3 Wednesday Thursday

4 Thursday Friday

5 Friday Saturday

6 Saturday Sunday

7 Sunday Monday

Let E be the event of having 52 Mondays and 53 Sundays in the year

Total number of possible cases is n=7

Number of favourable cases is m=1 (Saturday and Sunday is the only favourable case)

$$P(E)=\frac{M}{N}=\frac{1}{7}$$

2) A class consists of 10 Boys and 6 girls. If a Committee of 3 is chosen at random from the class, find the probability that

(i) 3 Bays are selected

(ii) Exactly 2 girls are selected

Sol: Total number of students = 16

n(S)= number of ways of choosing 3 from 16 = 16_{C_3}

(i)Suppose 3 boys are selected

This can be done in n(E)=10_{C_3} ways.

P(E) =Probability 3 boys are selected $=\frac{10_{C_3}}{16_{C_3}}$ =0.2143

(ii) Suppose exactly 2 girls are selected then

n(E)= $6_{C_2} \times 10_{C_1}$

$$\frac{n(\mathrm{E})}{n(\mathrm{S})}=\frac{6_{C_2}\times 10_{C_1}}{16_{C_3}}=0.2678$$

(3) Two Cords at are selected at random numbered 1 to 10. Find the probability that the Sum is even if

(i)2 cards are drawn together

(ii) 2 Cards drawn one after other with replacement

Sol: Suppose 2 Cards are drawn at a time

Number of ways of drawing 2 cards at a time from 10 cards = 10_{C_2} = 45 ways.

For the Sum on both the Cards to be even both the Cards should be even number on both the cards should be odd number

2 Even number cards can be chosen from 5 even number cards = 5_{C_2} = 45 ways.

Total number of favourable outcomes= 10+10=20

(ii) Suppose the 2 Cards are chosen one after another with replacement, This can be done in 10x10 = 100 ways

For the Sum to be even both the Cards must be even or both the Cards must be odd.

Number of ways selecting 2 even Cards= $5_{C_1} \times 5_{C_1}$ = 25 ways

Similarly

Number of ways of selecting 2 odd Cards= $5_{C_1} \times 5_{C_1}$ = 25 ways

Required probability $\frac{25+25}{100} = \frac{50}{100} = \frac{1}{2}$

(4)Two cards are selected at random from 10 each numbered 1 to 10. Find the probability that the Sum is odd

(i)If 2 cards are drawn

(ii)2 Cards are drawn one after another with replacement

(iii) 2 cards are draws one after another without replacement

Sol:

(i)2 Cards can be drawn at a time from 10 cards in 10_{C_2} = 45 ways

Let E_1 denote the event of 2 cards are such that the Sum is odd.

We must have one card even and another odd.

Number of ways of doing it= $5_{C_1} \times 5_{C_1}$ = 25

Required probability = 25/45 = 5/9

(ii) Let E_2 = The Sum is even when two cards are drawn one after another with replacement. The number of favourable cases= 50

∴The number of ways in which 2 cards can be drawn one after another with replacement

$=10_{C_1} \times 10_{C_1} = 100$

Required probability =50/100 = 1/2

The Number of favourable cases= 50

The number of cases that the two cards can be drawn one after another without replacement

$=10_{C_1} \times 9_{C_1} = 90$

Required probability= 50/90 = 5/9

(5) 5 digit numbers are formed with 0,1,2,3,4 (not allowing a digit being repeated in any number. Find the probability of getting 2 in the l0's place and 0 in the units place always.

Sol:

The total number of 5 digit no's using the digits 0,1,2,3,4 is n= 5!-4!= 96

Let E be the event of getting a number having 2 in 10's place and 0 in the units place

so that number of favourable = $3 \times 2 \times 1 \times 1 \times 1 = 6$

P(E)= 6 /96=1/16

(6) Out of 15 items 4 are not in good condition. 4 are selected at random. Find the probability that

(i) All are not good

(ii) 2 are not good

Sol:

Total no of items = 15

Number of ways of picking 4 items is 15_{C_4}

Suppose 4 items are chosen which are not good

number of ways selecting = $\frac{4_{C_4}}{15_{C_4}} = 1/1365$

(ii) Suppose 2 items are not good

Number of ways of selecting of 2 bad items= 4_{C_2}

$\therefore$Probability of getting 2 items= $\frac{^4C_2}{^{15}C_4}$ =2/455

1.8 SUMMARY

Probability is a fundamental concept in mathematics that quantifies uncertainty and randomness. It is defined as the measure of the likelihood of an event occurring, expressed as a number between 0 and 1. The axioms of probability, established by Kolmogorov, provide a formal foundation, including non-negativity, normalization, and additivity. Several elementary theorems, such as the addition and multiplication rules, help in solving probability problems efficiently. Worked-out examples illustrate these concepts in practice, demonstrating their application in real-world scenarios. Understanding probability and its theorems is essential for statistical analysis, decision-making, and various scientific disciplines.

1.9 TECHNICAL TERMS

Probability

Sample Space

Random Experiment

Event

Conditional Probability

Independent Events

Kolmogorov Axioms

Bayes' Theorem

1.10 SELF ASSESSMENT QUESTIONS

Essay questions:

1. Define probability and give its mathematical formula.
2. What are mutually exclusive and exhaustive events?
3. State and explain Kolmogorov's three axioms of probability.
4. What is the difference between independent and dependent events?
5. State Bayes' theorem and its significance in probability.

Short Questions:

1. Explain the classical, frequentist, and Bayesian definitions of probability with examples.
2. Prove and explain the addition and multiplication theorems of probability.
3. Discuss the Law of Large Numbers and its applications in probability and statistics.
4. Solve a real-world problem using Bayes' theorem, explaining each step in detail.

5 What is the Central Limit Theorem (CLT)? Discuss its importance and provide an example.

1.11 FURTHER READINGS

1 "A First Course in Probability" – Sheldon Ross

2 "Introduction to Probability" – Dimitri P. Bertsekas & John N. Tsitsiklis

3 "Probability and Statistics" – Morris H. DeGroot & Mark J. Schervish

4 "Probability and Random Processes" – Geoffrey Grimmett & David Stirzaker

5 "Probability: Theory and Examples" – Rick Durrett

Dr. A.J.V. Radhika

LESSON-2

CONDITIONAL PROBABILITY

OBJECTIVES:

After going through this lesson, you will be able to

- Understand how the probability of an event changes when another event is known to have occurred.
- Extends to multiple independent events:
- Used for updating probabilities based on new evidence, essential in statistics, AI, and machine learning.

STRUCTURE OF THE LESSION:

2.1 Introduction

2.2 Definition

2.3 Multiplication theorem of probability.

2.4 Independent Events

2.5 Multiplication theorem of probability for independent events

2.6 Compound event : Worked out examples.

2.7 Bayes theorem

2.8 Worked out problems.

2.9 Summary

2.10 Technical Terms

2.11 Self Assessment Questions

2.12 Further Readings

2.1 INTRODUCTION

The concept of probability calculation changes as situation occurred such as to modify the probability of an event because something new is known. This idea leads to the basic concept conditional Probability, which is the probability of an event occurring given that another event has already occurred.

2.2 DEFINITION

The conditional probability of the occurrence of an event B given that an event A is known to have already occurred is denoted by P(B/A), where A and B are the events associated the same sample space.

Ex: Let us consider a random experiment of drawing a card frame a pack of cards. Then the probability of happening of the event A: “The card drawn is a king" given by

$$\mathrm{P(A)}\frac{4}{52}=\frac{1}{13}$$

Suppose that a card is drawn and we are informed that the drawn card is red. How does this information affect the likelihood of the event A?

If the event B: "The card drawn is red "has happened, the event 'Black card’ is not possible. Hence the probability of the event A must be computed relative to the new sample space ‘B’ which consists of 26 sample Points (red colour only) i.e, n(B) = 26 .

Among these 26 red cards, there are two (red) kings. So that $P(A/\mathrm{B})=\frac{n(A\cap B)}{n(\mathrm{B})}=\frac{2}{26}=\frac{1}{13}$

2.3 MULTIPLICATION THEOREM OF PROBABILITY.

Theorem: For two events A and B,

$$\begin{aligned} P(A\cap \mathrm{B}) &= P(A).\mathrm{P(B/A)}, \quad \mathrm{P(A)}>0 \\ &= P(B).P(A/\mathrm{B}), \quad \mathrm{P(B)}>0 \end{aligned}$$

Where P(B /A) represents conditional probability of occurrence of B when the event A has already happened and P(A/B) is the conditional probability of happening of A, given that B has already happened.

Proof: In the usual, notations, we have

$$P(A)=\frac{n(A)}{n(S)}\ ,\quad P(\mathrm{B})=\frac{n(\mathrm{B})}{n(S)}\ \&\ P(\mathrm{A}\cap\mathrm{B})=\frac{n(\mathrm{A}\cap\mathrm{B})}{n(S)} \qquad \longrightarrow(1)$$

For the conditional event A/B, the favourable outcomes must be one of the sample points of B. i.e, for the event A/B, the sample space is B and out of the n(B) sample points, $n(A\cap B)$ pertain to the occurrence of the event A.

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

Rewriting (from (1)) ,we get

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

$$= P(B).\,P(A/\mathrm{B}) \qquad \longrightarrow(2)$$

Similarly, (from(1)), we get

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

$$= P(\mathrm{A}).\,P(\mathrm{B}/\,\mathrm{A}) \qquad \longrightarrow(3)$$

From (2) and (3), we get the result.

Thus we have proved that "The probability of the simultaneous occurrence of two events A and B is equal to the product of the probability of one of these events and the conditional probability of the other, given that the first one has occurred".

2.4 INDEPENDENT EVENTS

Two (or) more events are said to be independent, if the happening (or) non-happening of any one of them does not, in any way, affect the happening of others.

Consider the experiment of throwing two dice, say die 1 and die 2. It is obvious that the occurring of a certain number of dots on the die 1 has nothing to do with a similar event for the die 2. The two are quite independent of each other.

Similarly, if we drawn two cards from a pack of cards in succession, then the results of the two draws are independent if the cards are drawn with replacement (i.e, if the first card is drawn is placed back in the pack before drawing the second card) and the results of the two draws are not independent if the cards are drawn without replacement.

Definition: An event A is said to be independent of another event B, if the conditional probability of A given B, i.e P(A/B) is equal to the unconditional probability of B, i.e, if P(A/B) = P(A).

It may be noted that the above definition is meaningful only when P(A/B) is defined, i.e if P(B)≠0 similarly, an event B is said to be independent of event A, if P(B/A) = P(B); P(A)≠0.

Theorem: If the events A and B are such that P(A) ≠ 0 , P/B) ≠ 0 and A is independent of B, then B is independent of A.

Proof: Since the event A is independent of B, we have

P(A/B) = P(A)

$$\frac{P(\mathrm{A}\cap\mathrm{B})}{\mathrm{P(B)}} = P(A) \implies P(\mathrm{A}\cap\mathrm{B}) = P(A)\ \mathrm{P(B)}$$

$$\therefore \frac{P(\mathrm{B}\cap A)}{\mathrm{P(A)}} = P(\mathrm{B}) \quad [\therefore \mathrm{P(A)}\neq 0\ \&\ (A\cap B) = P(\mathrm{B}\cap A)\,]$$

$$\implies P(\mathrm{B}/\,\mathrm{A}) = \mathrm{P(B)}$$

$\Rightarrow$B is independent of A.

2.5 MULTIPLICATION THEOREM OF PROBABILITY FOR INDEPENDENT EVENTS

Theorem: If A and B are two events with positive probabilities($P(A)\neq0$, $P(B)\neq0$), then A and B are independent if and only if $P(A\cap B)=P(A).P(B)$

Proof: We have

$P(A\cap B)=P(A).P(B/A)=P(B)P(A/B)$; $\rightarrow(1)$ $P(A)\neq0$, $P(B)\neq0$

If A and B are independent, i.e , A is independent of B and B is independent of A, then we have $P(A/B)=P(A)$ & $P(B/A)=P(B)$ $\rightarrow(2)$

From (1) and (2)

$$\frac{P(A\cap B)}{P(B)}=P(A)\Rightarrow P(A/B)=P(A)$$

$$\frac{P(A\cap B)}{P(A)}=P(B)\Rightarrow P(B/A)=P(B)$$

This implies that A and B are independent events. Hence, for independent events A and B, the probability that both of these occur simultaneously is the product of their respective probabilities. This rule is known as the multiplication rule of probability.

2.6 COMPOUND EVENT: WORKED OUT EXAMPLES.

When two or more events occur in conjunction with each other, their joint occurrence is called compound event.

Example: If two balls are drawn from a bag containing 4 green, 6 black, 7 white balls. The event of drawing two green balls or two white balls is a compound event.

1)Two marbles are drawn in succession from box containing. 10 red, 30 white , 20 blue and 15 orange marbles with replacement being made after each row. Find the probability that

(i) Both are white

(ii)First is red and second is white.

Sol: Total number of marbles=75

(i) Let E_1 be the event of first drawn marble is white then $P(E_1)=\frac{30}{75}$

Let E_2 be the event of second drawn marble is also white then $P(E_2/E_1)=\frac{30}{75}$

$\therefore$The probability of both marbles are white with replacement $P(E_1\cap E_2)=P(E_1).P(E_2/E_1)$

$$=\frac{30}{75}\cdot\frac{30}{75}=\frac{4}{25}$$

(ii) Let E_1 be the event of first drawn marble is red. Then $P(E_1)=\frac{30}{75}$

Let E_2 be the event the of second drawn marble is also white then $P(E_2/E_1)=\frac{30}{75}$

∴The probability that the first marble is red and second marble is white.

$P(E_1 \cap E_2) = P(E_1).P(E_2/E_1)$

$$=\frac{10}{75}\cdot\frac{30}{75}=\frac{4}{75}$$

2)Determine

(i)P(B/A) (or) B given A

(ii) P(A) given B compliment P(A/B^C)

Sol: If A and B are events with

P(A)=1/3, P(B)=1/4, $P(A\cup B)=1/2$

P(A)=1/3, P(B)=1/4, $P(A\cup B)=1/4$

(i)$P(B/A)=\frac{P(A\cap B)}{P(A)}$

$$=\frac{P(A)+P(B)-P(A\cup B)}{P(A)}$$

$P(B/A)=\frac{1/3+1/4-1/2}{1/3}=\frac{1/12}{1/3}=1/4$

(ii)$P(A/B^C)=\frac{P(A\cap B^C)}{P(B^C)}$ =P(A)- $P(A\cap B)$ =1/3-1/12=1/4

$P(B^C)$=1-P(B)=1-1/4=3/4

$P(A/B^C)=\frac{1/4}{3/4}=1/3$

3) A,B,C are aiming to shoot balloon. 'A' will succeed four times out of 5 attempts. The chance of B to shoot the balloon is 3 out of 4. That of 'C' is 2 out of 3. If the three aim the balloon simultaneously, then find the probability that atleast two of them hit the balloon.

Sol: P(A)=4/5

P(B)=3/4

P(C)=2/3

∴ The probability of A,B,C not hitting the target respectively are

$P(\overline{A})$ =1-P(A)=1-4/5=1/5

$P(\overline{B})$ =1-P(B)=1-3/4=1/4

$P(\overline{C})$ =1-P(C)=1-2/3=1/3

Now the probability that exactly 2 will hit the balloon is

$= P(A \cap B \cap \overline{C}) + P(A \cap \overline{B} \cap C) + P(\overline{A} \cap B \cap C)$

=P(A).P(B). $P(\overline{C})$ +P(A). $P(\overline{B})$ .P(C)+ $P(\overline{A})$ .P(B).P(C)

=4/5.3/4.1/3+4/5.1/4.2/3+1/5.3/4.2/3=13/30

The probability of all will hit the balloon

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

=4/5.3/4.2/3 = 2/5

The probability that atleast two of them will hit the target=13/30+2/5=5/6

4) Three machines I, II, III produce 40%, 30%, 30% of the total number of items of factory. The percentages of detective items of this machines are 4%, 2%, 3%. If a one item is selected in a random. Find the probability that the item is defective.

Sol: Let A,B,C are be the events that the machines I,II, and III be chosen respectively. And 'D' be the event which denotes the defective item by data.

P(A)=40/100

P(B)=30/100

P(C)=30/100

P(D/A)=4/100

P(D/B)=2/100

P(D/C)=3/100

∴ The probability that the selected item at random is defective is

P(D)=P(A).P(D/A)+P(B).P(D/B)+P(C).P(D/C)

=(40/100)(4/100)+(30/100)(2/100)+(30/100)(3/100)

=31/1000

5) 2 dice are thrown, Let A be the event that the Sum of the points the faces is 9. Let B be the event that atleast one number is 6. Find

(i) $P(A\cap B)$ **(ii)** $P(A\cup B)$ **(iii)** $P(A^C\cup B^C)$

Sol: There are 36 sample outcomes when 2 dice are thrown.

The event A= The sum as 9 occurs in the following way.

A={(3,6)(4,5)(5,4)(6.3)}

P(A)=4/36

The event B atleast 1 number is 6 occurs in the following way

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

P(B)=11/36

Now $(A\cap B)$ ={(3,6)(6,3)}

(i)$P(A\cap B)$ =2/36=1/18

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

=4/36+11/36-1/18

=13/36

(iii) $P(A^C\cup B^C)$ = $P((A\cap B)^C)$

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

=1-1/18

=17/18

2.7 S BAYE'S Theorem.

$E_1, E_2,\ldots..E_n$ are mutually exclusive and exhaustive events such that $P(E_i) > 0$ (i=1,2,…n) in a sample space S and A is any other event in S intersecting with every E_i (i.e A can only occur in combination with any one of the events $E_1, E_2,\ldots..E_n$) Such that P(A)>0. If E_i is any of the events of $E_1, E_2,\ldots..E_n$ where $P(E_1)$, $P(E_2)$, $P(E_n)$ and $P(A/E_1)$, $P(A/E_2)$,….. $P(A/E_n)$ are known then

$$P(E_k/A)=\frac{\mathbf{P(E_k)P(A/E_k)}}{\mathbf{P(E_1)P(A/E_1)+P(E_2)P(A/E_2)+\ldots\ldots+P(E_n)P(A/E_n)}}$$

Proof:

$E_1, E_2, \ldots E_n$ are n events such that $P(E_i)>0$ and $E_i \cap E_j = \phi$ for $i \neq j$ where i,j=1,2,…n also $E_1, E_2, \ldots E_n$ are exhaustive events of S and A is any other event of S where P(A)>0

$S = E_1 \cup E_2 \cup \ldots \cup E_n$

$$A = A \cap S = A \cap (E_1 \cup E_2 \cup \ldots \cup E_n)$$

$$= (A \cap E_1) \cup (A \cap E_2) \cup, \ldots \cup (A \cap E_n) \qquad \rightarrow (1)$$

Here $(A \cap E_1), (A \cap E_2), \ldots (A \cap E_n)$ are mutually exclusive events then

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

$$= \frac{P(E_k \cap A)}{P[(A \cap E_1) \cup (A \cap E_2) \cup \ldots \cup (A \cap E_n)]} \qquad [\text{ From (1) }]$$

$$= \frac{P(E_k \cap A)}{P[(A \cap E_1) + (A \cap E_2) + \ldots + (A \cap E_n)]}$$

By multiplication theorem / Conditional probability

$$P\left(E_k / A\right) = = \frac{P\left(E_k\right) P\left(A / E_k\right)}{P\left(E_1\right) P\left(A / E_1\right) + P\left(E_2\right) P\left(A / E_2\right) + \ldots\ldots + P\left(E_n\right) P\left(A / E_n\right)}$$

2.8 Worked out Problems

Ex (1) In a certain college 25% of boys and 10% of girls are studying mathematics. The girls constitute 60% of the student body.

(a) What is the probability that mathematics is being studied

(b) If a student is selected at random and is found to be studying mathematics, find the probability that the student is girl.

(c)Student is boy.

Sol: Given $P(\text{Boys}) = \frac{40}{100} = \frac{2}{5}$

$P(\text{Girls}) = \frac{60}{100} = \frac{3}{5}$

Probability that mathematics is studied given that the student is boy

$\therefore P(M/B)=\frac{25}{100}=\frac{1}{4}$

Probability that the mathematics is studied given that the student is a girl

$\therefore P(M/G)=\frac{10}{100}=\frac{1}{10}$

a) Probability that the student studied mathematics

$P(M)=P(B)\,P(M/B)+P(G)\,P(M/G)$

$=\frac{2}{5}\times\frac{1}{4}+\frac{3}{5}\times\frac{1}{10}$

$=\frac{1}{10}+\frac{3}{50}=\frac{8}{50}$

$=\frac{4}{25}$

b)By Bayes theorem probability of mathematics student is a girl

$P(G/M)=\frac{P(G)P(M/G)}{P(B)P(M/B)+P(G)P(M/G)}$

$=\frac{\frac{3}{5}\times\frac{1}{10}}{\frac{4}{25}}$

$=\frac{3}{8}$

c) By Bayes theorem probability of mathematics student is a boy

$P(B/M)=\frac{P(M)P(M/B)}{P(B)P(M/B)+P(G)P(M/G)}$

$=\frac{\frac{2}{5}.\frac{1}{4}}{\frac{4}{25}}=\frac{10}{16}=\frac{5}{8}$

2) **In a Bolt factory machines A, B, C manufacture 20%, 30% and 50% of the total of their output and 6%,3%, and 2% care defectives. A bolt is drawn (from) at Random and found to be defective. Find the probability that it is manufactured from (i) machine A (ii) machine B (iii) Machine C**

Sol: Let P(A), P(B), P(C) be the probabilities of the events that the bolts are manufactured by machines A,B,C respectively. Then P(A)=$\frac{20}{100}$

P(B)=$\frac{30}{100}$ and P(C)=$\frac{50}{100}$

Let D be the defective of the bolt, then

P(D/A)=$\frac{6}{100}$

P(D/B)=$\frac{3}{100}$

P(D/C)=$\frac{2}{100}$

(i) If a bolt is defective, then the probability it is from machine A

$$P(A/D)=\frac{P(A)P(D/A)}{P(A)P(D/A)+P(B)P(D/B)+P(C)P(D/C)}$$

$$=\frac{\frac{20}{100}\times\frac{6}{100}}{\frac{20}{100}\times\frac{6}{100}+\frac{30}{100}\times\frac{3}{100}+\frac{50}{100}\times\frac{2}{100}}=\frac{12}{31}$$

(ii) If a bolt is defective, then the probability it is from machine B

$$P(B/D)=\frac{P(B)P(D/B)}{P(A)P(D/A)+P(B)P(D/B)+P(C)P(D/C)}$$

$$=\frac{\frac{30}{100}\times\frac{3}{100}}{\frac{20}{100}\times\frac{6}{100}+\frac{30}{100}\times\frac{3}{100}+\frac{50}{100}\times\frac{2}{100}}$$

$$=\frac{9}{31}$$

(iii) If a bolt is defective, then the probability it is from machine C

$$P(C/D)=\frac{P(C)P(D/C)}{P(A)P(D/A)+P(B)P(D/B)+P(C)P(D/C)}$$

$$=\frac{\frac{50}{100}\times\frac{2}{100}}{\frac{20}{100}\times\frac{6}{100}+\frac{30}{100}\times\frac{3}{100}+\frac{50}{100}\times\frac{2}{100}}$$

$$=\frac{10}{31}$$

(3) A Businessman goes to hotels x, y, z 20% , 50% , 30% of the time respectively. It is known that 5%, 4%, 8% of the rooms in x, y, z hotels have faulty plumbing. What is the probability that Businessman's having faulty plumbing is assigned to hotel z.

Sol: Let P(x), P(y), P(z) be the probabilities of the events of going to hotels x, y, z

$$P(x)=\frac{20}{100}\ ,\ P(\mathrm{y})=\frac{50}{100}\ ,\ P(\mathrm{z})=\frac{30}{100}$$

Let D be the faulty plumbing

(i) Probability that Businessman's having faulty plumbing is assigned to hotel z is

$$P(\mathrm{z}/D)=\frac{P(z)P(D/\mathrm{z})}{P(\mathrm{x})P(D/\mathrm{x})+P(\mathrm{y})P(D/\mathrm{y})+P(\mathrm{z})P(\mathrm{D}/\mathrm{z})}$$

$$=\frac{\frac{30}{100}\times\frac{8}{100}}{\frac{20}{100}\times\frac{5}{100}+\frac{50}{100}\times\frac{4}{100}+\frac{30}{100}\times\frac{8}{100}}$$

$$=\frac{4}{9}$$

4) If the 3 men, the chances that the politician, a businessmen or an academician will be appointed as a Vice Chancellor of a university are 0.5, 0.3, 0.2 respectively. Probability that research is promoted by these three persons if they are appointed as VC are 0.3, 0.7, 0.8 respectively

(i)Determine the probability that research is promoted

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

Sol: Let A , B, C be the events that a politician businessmen or an academician will be appointed as VC of the three men . Then

P(A)=0.5

P(B)=0.3

P(C)=0.2

The probabilities that research is promoted if they are appointed as VC's are

P(R/A)= 0.3,

P(R/B)= 0.7,

And P(R/C)=0.8

(i)The probability that the research is promoted

=P(A).P(R/A)+P(B).P(R/B)+P(C).P(R/C)

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

=0.52

(ii) The probability that research is promoted when the VC is an academician

$$P(\mathrm{C}/\mathrm{R}) = \frac{P(C).P(\mathrm{R}/\mathrm{C})}{P(\mathrm{C}).P(\mathrm{R}/\mathrm{C}) + P(\mathrm{B}).P(\mathrm{R}/\mathrm{B}) + P(\mathrm{A}).P(\mathrm{R}/\mathrm{A})}$$

$$= \frac{0.16}{0.15 + 0.21 + 0.16}$$

$$= \frac{4}{13}$$

=0.30769

2.9 SUMMARY

Conditional probability helps determine the likelihood of an event occurring given that another event has already happened, forming the basis for the Multiplication Theorem of Probability, which states that P(A∩B)=P(A)P(B|A) . When events are independent, their occurrence does not affect each other, simplifying the multiplication rule to P(A∩B)=P(A)P(B). Compound events involve multiple outcomes and can be either dependent or independent. Bayes' Theorem provides a way to update probabilities based on new evidence, making it crucial in decision-making, statistics, and machine learning. These concepts form the foundation of probability theory and real-world applications.

2.10 TECHNICAL TERMS

1. Conditional Probability
2. Multiplication Theorem
3. Independent Events
4. Joint Probability

5. Compound Events
6. Bayes' Theorem
7. Prior Probability
8. Posterior Probability

2.11 SELF ASSESSMENT QUESTIONS

Short questions:

1. Define conditional probability and provide its formula.
2. What is the multiplication theorem of probability?
3. How do you determine if two events are independent?
4. Explain the concept of a compound event with an example.
5. State Bayes' Theorem and its significance in probability.

Essay Short Questions:

1. **Derive the formula for conditional probability with a real-world example.**
2. Explain the multiplication theorem for both dependent and independent events.
3. Discuss independent events with examples and prove the multiplication rule for them.
4. What are compound events? Differentiate between independent and dependent compound events with examples.
5. Explain Bayes' Theorem in detail with an application in real-life scenarios (e.g., medical testing or spam filtering).

2.12 FURTHER READINGS

1 "A First Course in Probability" – Sheldon Ross

2 "Introduction to Probability" – Dimitri P. Bertsekas & John N. Tsitsiklis

3 "Probability and Statistics" – Morris H. DeGroot & Mark J. Schervish

4 "Probability and Random Processes" – Geoffrey Grimmett & David Stirzaker

5 "Probability: Theory and Examples" – Rick Durrett

Dr. A.J.V. Radhika

Lesson-3

RANDOM VARIABLES

OBJECTIVES:

After going through this lesson, you will be able to

- Understand Random Variables – Learn the definition and types of random variables (discrete and continuous).
- Explore Discrete Probability Distributions – Study how probabilities are assigned to discrete random variables.
- Analyze Key Parameters – Understand the significance of parameters like mean, variance, and standard deviation in probability distributions.
- Apply Theoretical Concepts – Learn important probability theorems and their applications in solving real-world problems.
- Develop Problem-Solving Skills – Work through solved examples to apply probability concepts in practical scenarios.

STRUCTURE OF THE LESSION:

3.1 Introduction

3.2 Random variable

3.3 Discrete probability distribution

3.4 Parameters

3.5 Some theorems.

3.6 Worked out problems.

3.7 Summary
3.8 Technical Terms
3.9 Self-Assessment Questions

3.10 Further Readings

3.1 INTRODUCTION

Suppose S is the Sample space of some experiment we know that outcomes of the experiment are the elements of the Sample space S and they need not be numbers. Sometimes (we wish) to assign a specify number to each outcome.

Example: The number of heads in tossing 2 coins or 3 coins such assignment is called a Random Variable.

In the above example we may consider the Random Variable which is the number of heads.

Outcome: HH HT TH TT

Variables: 2 1 1 0

3.2 RANDOM VARIABLE

A Real Value X whose value is determined by the outcome of the Random Experiment is called a Random Variable.

3.2.1 Types of Random Variable

Random Variables is of 2 types:

(a) Discrete Random Variable.

(b) Continuous Random Variable

3.2.2 Discrete Random Variable:

Random Variable X which can take only a finite number of discrete values in an interval of domain is called Discrete Random Variables.

In other words,

If the Random Variables takes if the values only on the set {0,1,2,…n} is called a Discrete Random Variable.

Eg: If a coin is tossed X(H) = 1 if Head occurs.

X(H)= 0 if tail occurs.

3.2.3 Continuous Random Variable:

Random Variable X which can take Values Continuously i.e which takes all possible values in a given interval is called a Continuous Random Variable.

Eg: The Height, age, weight, temperature

Probability function of a Discrete Random Variable:

If for a Discrete Random Variable X, the real valued function P(X) is such that P(X= x) =P(x) then P(x) is called probability function or Probability density function of a discrete Random Variable X.

3.2.4 Properties of a probability function:

If P(X) is a probability function of a Random Variable X, then it is possess the following properties

(i) $P(x \geq 0) \ \forall \ x$

(ii) $\sum P(X) = 1$

(iii)P(X) can not be negative for any value of X.

3.2.5 Probability Distribution function

Definition: Let X be a Random Variable. Then the probability distribution function associated with x is defined as the probability that the outcome of an experiment will be one of the outcomes for which $X(S) \leq x$ ($\therefore$ S $\longrightarrow$ sample space and x $\longrightarrow$variable)

$F_x(x) = P(X \leq x) : -\infty \leq x \leq \infty$ ($\therefore F_x \longrightarrow$ Distribution function)is called the Distribution function of X.

Properties of Distribution function:

(1) If F is the distribution function of a Random Variable x and if a < b then

(a)P(a<x<b)=F(b)=F(a)

(b) $P(a \leq x \leq b) = P(x = a) + [F(b) - F(a)]$

(c) P(a<x<b)= [F(b) - F(a)] -P(x=b)

(d) $P(a \leq x \leq b) = [(F(b) - F(a))] - P(x = b) + P(x = a)$

(2) $F(-\infty) = \lim_{x \to -\infty} F(x) = 0$

3.3 DISCRETE PROBABILITY DISTRIBUTION (PROBABILITY MASS FUNCTION [PMF])

Suppose X is a Discrete Random Variable taking atmost infinite number of values $x_1, x_{2...} x_n$ the probability of each possible outcome x_i ,we associated a number $P_i = P(x = x_{i)} = P(x_i)$ is called the probability of X_i, i = 1,2…n must satisfy the following conditions

(i) $P(x_i \geq 0)$ $\forall$ Values of i

(ii) $\sum P(x_i) = 1$, i=1, 2,….is called the PMF of Random variable X and the set $P(x_i)$ is called Discrete Probability Distribution of the Discrete Random Variable.

3.4 PARAMETERS

Expectation, Mean, Variance and Standard Deviation of a discrete probability distribution:

(1) Expectation:

$$E(x) = \sum x\ p(x)$$

(2) Mean:

$$\mu = \sum x\ p(x)$$

(3) Variance:

$$V(x) = \sigma^2 = \sum (x_i - \mu)^2 \; p(x) = E(\mathrm{X}^2) - [\mathrm{E(X)}]^2 = E(\mathrm{X}^2) - \mu^2$$

(4) Standard deviation:

$$\mathrm{SD} = \sqrt{V(x)}$$

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

3.5 SOME THEOREMS

Theorem 1:

Statement: If x is a Random variable, k is a constant, then E(x+k)=E(x)+k

Proof: By definition

$$\mathrm{E(x+k)} = \sum_{i=1}^{n} (x+k)\,p(x)$$

$$= \sum_{i=1}^{n} x\; p(x) + \sum_{i=1}^{n} k\; p(x)$$

$$= E(\mathrm{x}) + k \sum_{i=1}^{n} p(x)$$

$$= E(\mathrm{x+k}) = E(x) + k \qquad \left[\therefore \sum p(x) = 1\right]$$

Hence Proved

Theorem 2:

Statement: If x is a Random variable, a, b are constants that E(ax+b)=aE(x)+b

Proof:

By definition

$$\mathrm{E(ax+b)} = \sum_{i=1}^{n} (ax+b)\,p(x)$$

$$= \sum_{i=1}^{n} ax\; p(x) + \sum_{i=1}^{n} b\; p(x)$$

$$= a \sum_{i=1}^{n} x\; p(x) + b \sum_{i=1}^{n} p(x)$$

$$\mathrm{E(ax+b)=a\,E(x)+b} \qquad \left[\therefore \sum p(x) = 1\right]$$

Hence Proved.

Theorem 3:

<u>Statement:</u> If X and Y are 2 Random Variables then E(x+y)=E(x)+E(y) provided E(X) and E(Y) exists.

<u>Proof:</u> Let X assume the values $x_1, x_2, \ldots x_n$ and Y assumes $y_1, y_2, \ldots y_m$ then by definition

$$E(X)=\sum_{i=1}^{n} x_i p_i$$

$$E(Y)=\sum_{j=1}^{m} y_j p_j$$

Let $p_{ij} = (X = x_i \cap Y = y_j) = p(x_i, y_j)$

(This is called joint probability function of x and y)

The sum X+Y is also called a random variable which can take $m \times n$ values (x_i+y_j), i=1,2…n

And j=1,2,…m

∴By definition

$$E(X+Y)=\sum_{i=1}^{n}\sum_{j=1}^{m} p_{ij}(x_i + y_j)$$

$$=\sum_{i=1}^{n}\sum_{j=1}^{m} p_{ij}x_i + \sum_{i=1}^{n}\sum_{j=1}^{m} p_{ij}y_j$$

$$=\sum_{i=1}^{n}\left[x_i \sum_{j=1}^{m} p_{ij}\right] + \sum_{j=1}^{m}\left[y_j \sum_{i=1}^{n} p_{ij}\right]$$

$$=\sum_{i=1}^{n} x_i p_i + \sum_{j=1}^{m} y_j p_j = E(X)+E(Y)$$

Hence Proved.

Note:

1)E(X+Y+Z)=E(X)+E(Y)+E(Z)

2)E(AX+BY)=A E(X)+ B E(Y|) Where A and B are constants.

3) $E(X - \overline{X}) = 0$

3.6 WORKED OUT PROBLEMS

1) A random variable X has the following probability function

x	0	1	2	3	4	5	6	7

P(x)	0	k	2k	2k	3k	k^2	$2k^2$	$7k^2+k$

(i)Determine k

(ii)Evaluate P(x<6), P(x≥6), P(0<x<5) and P(0≤x≤4)

(iii)If P(x≤k>1/2) find the minimum value of k

(iv) Determine the distribution function of X

(v)Mean

(vi)Variance

Sol:

(i) We have $\sum_{i=0}^{7} P(x) = 1$

$0+k+2k+2k+3k+ k_2+2k^2+7k^2+k=1$

$10k^2+9k-1=0$

$10k^2+9k-1=0$

$10k^2+10k-k-1=0$

$10k(k+1)-1(k+1)=0$

$(10k-1)(k+1)=0$

$k=1/10, k=-1$

$\therefore k=1/10 \qquad \begin{pmatrix} \therefore P(x) \geq 0 \\ S\theta.k \neq -1 \end{pmatrix}$

(ii)

(a) p(x<6)=p(x=0)+p(x=1)+p(x=2)+p(x=3)+p(x=4)+p(x=5)

=0+k+2k+2k+3k+k

$=k^2+8k$

=1/100+8/10

=81/100

=0.81

(b)

P(x≥6)=1-P(x<6) ==1-81/100=19/100

(or)

p(x=6)+p(x=7)

$2k^2+7k^2+k$

$9k^2+k=9(1/100)+1/10$

$=19/100$

(c)

$p(0<x<5)=p(x=1)+p(x=2)+p(x=3)+p(x=4)$

$=k+2k+2k+3k$

$=8k$

$=8/10=0.8$

(d)

$p(0\leq x\leq 4)=P(x=0)+p(x=1)+p(x=2)+p(x=3)+p(x=4)$

$=0+k+2k+2k+3k$

$=8k=8/10=0.8$

(iii)The required minimum value of k is obtained as follows:

$P(x\leq 1)=p(x=0)+p(x=1)$

$=0+k$

$=1/10$

$=0.1$

$P(x\leq 2)=p(x\leq 1)+p(=2)$

$=0.1+2/10$

$=0.3$

$P(x\leq 3)=p(x\leq 2)+p(x=3)$

$=0.3+2/10$

$=0.5$

$P(x\leq 4)=p(x\leq 3)+p(x=4)$

$=0.5+3/10$

$=0.8>0.5=1/2$

$\therefore$The minimum value of 'k' for which $p(x\leq k>1/2)$ is k=4

(iv) The distribution of function x is

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

0	0
1	k=1/10
2	3k=3/10
3	5k=5/10
4	8k=8/10
5	$8k+k^2=81/100$
6	$81+3k^2=83/10$
7	$9k+10k^2=1$

(v) Mean= $\sum_{i=1}^{7} x\ p(x)$

$=0(0)+1(k)+2(2k)+3(2k)+4(3k)+5(k^2)+6(2k^2)+7(7\ k^2+k)$

$=k+4k+6k+12k+5k^2+12k^2+49k^2+7k$

$=66k^2+30k$

$=66(1/100)+30(1/10)$

$=0.66+3$

E(x)= μ =3.66

(vi)Variance

$V(x)=E(x^2)-(E(x))^2$

$=\sum x^2 p(x)-\mu^2$

$=[(0)^2(0)+(1)^2(k)+2^2(2k)+2^2(2k)+3^2(2k)+4^2(3k)+5^2(k^2)+6^2(2k^2)+7^2(7k^2+k)]-(3.66)^2$

$=[0+1(1/10)+4(4/10)+9(2/10)+16(3/10)+25(1/100)+36(4/100)+49(49(1/10)^2+1/10)]-(3.66)^2$

$=4.4+12.4-(3.66)^2=3.4044=\sigma^2$

Standard variation= $\sqrt{V(x)}=\sqrt{3.404}=1.845$

2) 1) A random variable X has the following probability function

x	**-3**	**-2**	**-1**	**0**	**1**	**2**	**3**
P(x)	**k**	**0.1**	**k**	**0.2**	**2k**	**0.4**	**2k**

Find (i)k (ii)Mean (iii)Variance

Sol:

(i) We have $\sum_{i=-3}^{3} P(X)=1$

$\Rightarrow$k+0.1+k+0.2+2k+0.4+2k=1

6k+0.7=1

6k=1-0.7

k=$\frac{1-0.7}{6}$=0.05

(ii) $\mu = \sum_{i=-3}^{3} X_i P(X_i)$

=(-3)(0.05)+(-2)(0.1)+(-1)(0.05)+0(0.2)+1(2(0.05))+2(0.4)+3(2(0.05)

E(x)= μ =0.8

(iii)Variance

$V(X) = E(X^2) - [E(X)]^2 = \sum x^2 p(x) - \mu^2$

=$(-3)^2$ (0.05)+$(-2)^2$(0.01)+(-1)(0.05)+0(0.02)+1(2(0.05))+2^2(0.4)+9(2(0.05))

=143/50

Theorem:

If x is discrete random variable then v(ax+b)=a^2v(x) where v(x) is variance of x and a, b are constants.

Proof:

Let y=ax+b $\longrightarrow$(1)

Then E(y)=E(ax+b)

=a E(x)+b $\longrightarrow$(2)

(1)-(2) given (y-E(y)=ax+b-aE(x)-b

(y-E(y))=a[x-E(x)]

Squaring and taking expectations on both sides on both sides

$E(y-E(y)^2)=a^2E[x-E(x)^2]$

$v(y)=a^2v(x)$ $[\therefore E(X-\mu)^2 = v(x)]$

$v(ax+b)=a^2v(x)$

case 1:

If b=0

Then $v(ax)=a^2x$

If a=0

Then v(b)=0

If a=1

Then v(ax+b)=b(x)

Note: If x and y are 2 independent random variables then v(x±y)=v(x)±v(y)

1)Let X denote the minimum of the 2 numbers that appear when a pair of fair dice is thrown once. Determine the

(i)Discrete probability distribution

(ii)Expectation

(iii)Variance

Sol: When 2 dice are thrown total number of outcomes= 6×6 =36

∴Sample space S it is as follows

$$S=\begin{pmatrix}(1,1),(1,2),\ (1,\ 3),(1,4),(1,5),(1,6)\\(2,1),(2,2),(2,3),(2,4),(2,5),(2,6)\\(3,1),(3,2)\ (3,3),(3,4),(3,5),\ (3,6)\\(4,1),(4,2),(4,3),(4,4),(4,5),(4,6)\\(5,1),(5,2),(5,3),(5,4),(5,5),(5,6)\\(6,1),(6,2),(6,3),(6,4),(6,5),(6,6)\end{pmatrix}$$

If the random variable X assigns the minimum of its number in S then the sample space S is

$$S=\begin{Bmatrix}1&1&1&1&1&1\\1&2&2&2&2&2\\1&2&3&3&3&3\\1&2&3&4&4&4\\1&2&3&4&5&5\\1&2&3&4&5&6\end{Bmatrix}$$

The minimum number could be {1,2,3,4,5,6}

For minimum 1, favourable cases are

{(1,1,)(1,2)(1,3)(1,4)(1,5)(1,6)(2,1)(3,1)(4,1)(5,1)(6,1)} //non repeated Eg {(2,1)but not(1,2)//

∴P(X=1)=$\frac{11}{36}$

Similarly

P(X=2)=$\frac{9}{36}$

$P(X=3)=\frac{7}{36}$

$P(X=4)=\frac{5}{36}$

$P(X=5)=\frac{3}{36}$

$P(X=6)=\frac{1}{36}$

(i) The probability function is

x	1	2	3	4	5	6
P(x)	$\frac{11}{36}$	$\frac{9}{36}$	$\frac{7}{36}$	$\frac{5}{36}$	$\frac{3}{36}$	$\frac{1}{36}$

(ii) $\sum_{i=1}^{6} E(X) = \sum_{i=1}^{6} X\ P(X)$

=11/36+18/36+21/36+20/36+15/36+6/36

=91/36

(iii)Variance:

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

$=1^2(11/36)+2^2(9/36)+3^2(7/36)+4^2(5/36)+5^2(3/36)+6^2(1/36)-(91/36)^2$

$=(11/36+36/36+63/36+80/36+75/36+36/36)-\frac{8281}{1296}$

$=\frac{301}{36}-\frac{8291}{1296}=\frac{-7679}{1296}=$ -5.925

2)Find the mean and variance of the uniform probability distribution given by f(x)=1/n. for x=1,2,3,…n.

Sol: The probability distribution is

x 1 2 3 …… n

f(x) $\frac{1}{n}$ $\frac{1}{n}$ $\frac{1}{n}$…….n

(i)Mean = $\mu = \sum x\ p(x)$

$$= 1\left(\frac{1}{n}\right)+2\left(\frac{1}{n}\right)+3\left(\frac{1}{n}\right)+.....+n\left(\frac{1}{n}\right)$$

$$= \frac{1}{n}(1+2+3+....+n)$$

$$= \frac{1}{n}.\frac{n(n+1)}{2}$$

$$= \frac{(n+1)}{2}$$

(ii)$V(x)=E(x)^2-[E(x)]^2$

$=1^2.1/n+2^2.1/n+….+n^2.1/n-\left(\frac{n+1}{2}\right)^2$

$=\frac{1}{n}(1^2+2^2+…..+ n^2)-\left(\frac{n+1}{2}\right)^2$

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

$=\frac{n^2-1}{12}$

3) A sample of 4 items is selected at random from a box containing 12 items of which 5 are defective. Find the expected number E of defective items.

Sol: Let x denote the number of defective items among 4 items drawn from 12 items.

Obviously, x can take the values 0,1,2,3, and 4

Number of good items=7

Number of defective items=5

P(x=0)=P(no defective)=$\frac{7c_4}{12c_4}=7/99$

P(x=1)=P(one defective and 3 good items)=$\frac{5c_1\times 7c_3}{12c_4}=35/99$

P(x=2)=P(Two defective and 2 good items)

$\frac{5c_2\times 7c_2}{12c_4}=42/99$

P(x=3)=P(3 defective and one good)

$$=\frac{5c_3 \times 7c_1}{12c_4}=14/99$$

P(x=4)=P(all are defective)

$$\frac{5c_4}{12c_4}=1/99$$

Discrete probability distribution is

x	0	1	2	3	4
P(x)	$\frac{7}{99}$	$\frac{35}{99}$	$\frac{42}{99}$	$\frac{14}{99}$	$\frac{1}{99}$

Expected number of defective items $=\sum x\ p(x)$

$$=0\left(\frac{7}{99}\right)+1\left(\frac{35}{99}\right)+2\left(\frac{42}{99}\right)+3\left(\frac{14}{99}\right)+4\left(\frac{1}{99}\right)$$

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

4) A fair die is tossed . Let random variable x denote the twice the number appearing on the die

(i) Write the probability distribution of x

(ii)Mean

(iii)Variance

Sol: Let x denote twice the number appearing on the face when a die is thrown.

Then x is a discrete random variable whose probability distribution is given by

(i)

$X=x_i$	2	4	6	8	10	12
$P(x_i)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

(ii)Mean=E(x)=$\frac{1}{6}=p_i\,x_i$

$$=2\times\frac{1}{6}+4\times\frac{1}{6}+6\times\frac{1}{6}+8\times\frac{1}{6}+10\times\frac{1}{6}+12\times\frac{1}{6}$$

$$=\frac{42}{6}=7$$

Now

$E(x^2)=\sum x_i^2 p(x_i)$

$$=2^2\times\frac{1}{6}+4^2\times\frac{1}{6}+6^2\times\frac{1}{6}+8^2\times\frac{1}{6}+10^2\times\frac{1}{6}+12^2\times\frac{1}{6}$$

$$=\frac{364}{6}$$

$$=60.67$$

(iii)Variance=$E(X^2)-[E(x)]^2$

$=60.67-(7)^2$

$=60.67-49$

$=11.67$

$$\text{stan}dard\ \ deviation=\sqrt{11.67}$$

$$=3.416138171$$

3.7 SUMMARY

This chapter introduces the concept of random variables, which are fundamental in probability theory and statistical analysis. A random variable can be either discrete or continuous, with a discrete probability distribution assigning probabilities to distinct outcomes. Important parameters, such as mean, variance, and standard deviation, help in understanding the characteristics of these distributions. Several key theorems provide a mathematical foundation for probability calculations and statistical inference. To reinforce these concepts, worked-out problems are included, demonstrating practical applications and problem-solving techniques in probability and statistics.

3.8 TECHNICAL TERMS

- Random Variable
- Discrete Probability Distribution
- Probability Mass Function (PMF)
- Expectation (Mean)
- Variance
- Standard Deviation
- Binomial Distribution
- Poisson Distribution

3.9 SELF-ASSESSMENT QUESTIONS

Short Questions:

1. Define a random variable and give an example.
2. What is a discrete probability distribution?
3. Explain the concept of expectation (mean) of a random variable.
4. What are the key parameters of a probability distribution?
5. State and explain any one theorem related to probability distributions.

Essay Questions:

1. Explain the difference between discrete and continuous random variables with examples.
2. Describe the properties and significance of a discrete probability distribution.
3. Derive and explain the formulas for expectation and variance of a discrete random variable.
4. Discuss important theorems related to probability distributions with proofs.
5. Solve a real-world problem using a binomial or Poisson probability distribution.

3.10 FURTHER READINGS

1. "An Introduction to Probability Theory and Its Applications, Volume 1" – William Feller, Wiley, 1968.
2. "Probability Theory: A Concise Course" – Y. A. Rozanov, Dover Publications, 1977.
3. "Probability Theory: An Introductory Course" – Iakov G. Sinai, Springer, 1992.
4. "Probability Theory: A Comprehensive Course" – Achim Klenke, Springer, 2020.
5. "Fat Chance: Probability from 0 to 1" – Benedict Gross, Joe Harris, Emily Riehl, Cambridge University Press, 2019.

Dr. A.J.V. Radhika

Lesson-4

DISCRETE DISTRIBUTION

OBJECTIVES:

After going through this lesson, you will be able to

- Understand Discrete Probability Distributions – Learn the fundamental concepts of discrete probability distributions and their applications.
- Explore Binomial Distribution – Study the binomial distribution, its probability mass function (PMF), and real-world applications.
- Analyze Constants of Binomial Distribution – Understand key parameters such as mean, variance, and standard deviation and their significance.
- Apply Binomial Distribution Through Worked-Out Problems – Solve numerical problems to enhance problem-solving skills.
- Learn Poisson Distribution – Understand the Poisson distribution as a limiting case of the binomial distribution and its importance in modeling rare events.
- Identify Constraints of Poisson Distribution – Study the conditions under which the Poisson distribution is applicable.
- Develop Problem-Solving Skills – Work on solved examples to gain practical understanding of binomial and Poisson distributions.

STRUCTURE OF THE LESSION:

4.1 Discrete distribution

4.2 binomial distribution

4.3 Constants of binomial distribution

4.4 Worked out problems

4.5 Poisson distribution

4.6 Constraints of Poisson distribution

4.7 Worked out Problems

4.8 Summary

4.9 Technical Terms

4.10 Self-Assessment Questions

4.11 Further Readings

4.1 DISCRETE DISTRIBUTION

Definition: A random variable X has a discrete distribution if its probability distribution is given by P(x)= $P(X)=\frac{1}{k}$ for X=$x_1, x_2, \ldots x_n$.

Random variable X is called Discrete distribution.

For example, x : 0 1

P(x) : $\frac{1}{2}$ $\frac{1}{2}$

4.2 BINOMIAL DISTRIBUTION

A Random variable X is said to follow a Binomial distribution if it assumes only a non-negative values and its probability mass function is given t

$$P(X=x)=\begin{cases} n c_x \; p^x \quad q^{n-x} & x=0,1,2,\ldots..n \; and \;\; q=1-p \\ 0 \quad otherwise \end{cases}$$

4.3 CONSTANTS OF BINOMIAL DISTRIBUTION

(i) Mean

$$\mu = E(x) = \sum_{x=1}^{n} x \; p(x)$$

$$= \sum_{x=1}^{n} x.\, n c_x p^x \quad q^{n-x}$$

$$= \sum_{x=1}^{n} x.\, \frac{n}{x}.\; n-1 c_{x-1} \; p^{x-1} \cdot \; p \; q^{n-x}$$

$$= np \sum_{x=0}^{n} \; n-1 c_{x-1} \; p^{x-1} \quad q^{n-x}$$

$$= np(q+p)^{n-1} \qquad \text{[From binomial distribution]}$$

$$= \; \mu = np \qquad (\therefore p+q=1)$$

(ii) Variance of Binomial distribution

$V(x)=E(X^2)-(E(X))^2$

Consider

$$E(X^2)= \sum_{x=0}^{n} x^2 p(x)$$

$$= \sum_{x=0}^{n} x^2 . n c_x p^x \quad q^{n-x}$$

$$= \sum_{x=0}^{n} [x(x-1)+x]\, n_{c_x} p^x \quad q^{n-x}$$

$$= \sum_{x=0}^{n} x(x-1)\, n_{c_x} p^x \quad q^{n-x} + \sum_{x=0}^{n} x\, n_{c_x} p^x \quad q^{n-x}$$

$$= \sum_{x=0}^{n} x(x-1) \frac{n}{x} \frac{n-1}{x-1}\; n-2_{c_{x-2}} p^{x-2} \quad q^{n-x} p^2 \quad q^{n-x} + np$$

$$= \sum_{x=0}^{n} \quad n(n-1)\, n-2_{c_{x-2}} p^{x-2} \quad p^2 \quad q^{n-x} + np$$

$$= n(n-1)p^2 \quad \sum_{x=0}^{n} \quad n-2_{c_{x-2}} p^{x-2} \quad q^{n-2} + np$$

$$= n(n-1)p^2 \quad (q+p)^{n-2} + np$$

$$E(\mathrm{X}^2) = n(n-1)p^2 + np$$

$$\mathrm{V(X)} = n(n-1)p^2 + np - (np)^2$$

$$= \mathrm{n^2p^2 - np^2 + np - n^2p^2}$$

=np(1-p)

V(x)=npq

$$SD = \sqrt{npq}$$

(iii) Moment generating function of Binomial distribution:

$$M_X(t) = E(e^{tx})$$

$$= \sum_{x=0}^{n} e^{tx} \quad .\mathrm{p(x)}$$

$$= \sum_{x=0}^{n} e^{tx}\, n_{c_x} p^x \quad q^{n-x}$$

$$= \sum_{x=0}^{n} n_{c_x} (pe^t)^x \quad q^{n-x}$$

$$M_X(t) = (q + pe^t)^n$$

(iv) Mode of Binomial distribution:

We know that

$$p(x) = n_{c_x} p^x \quad q^{n-x} \qquad \rightarrow (1)$$

$$p(x+1) = n_{c_{x+1}} p^{x+1} \quad q^{n-(x+1)} \qquad \rightarrow (2)$$

Equation $(2) \div (1)$

$$p(x) = n_{c_x} p^{x} \quad q^{n-x} \qquad \rightarrow (1)$$

$$\frac{n_{c_{x+1}} p^{x+1} \quad q^{n-(x+1)}}{n_{c_x} p^{x} \quad q^{n-x}} = \frac{p(x+1)}{p(x)}$$

$$p(x+1) = \frac{n-x}{x+1}\frac{p}{q}\, p(x)$$

4.4 WORKED OUT PROBLEMS

1)10 coins are tossed simultaneously. Find the probability of getting atleast

(i) 7 heads

(ii) 6 heads

Sol: p= probability of getting head=1/2

q=probability of not getting a head=1/2

The probability of getting X heads in a throw of 10 coins is

$$\text{P(X=x)= } p(x) = 10_{c_x} \left(\frac{1}{2}\right)^{x} \quad \left(\frac{1}{2}\right)^{n-x} \qquad x = 0,1,2,....10$$

(i)Probability of getting atleast 7 heads is given by

$$\text{P}(x \geq 7) = P(\text{X} = 7) + \text{P}(\text{X} = 8) + \text{P}(\text{X} = 9) + \text{P}(\text{X} = 10)$$

$$= 10_{c_7} \left(\frac{1}{2}\right)^{7} \quad \left(\frac{1}{2}\right)^{10-7} + 10_{c_8} \left(\frac{1}{2}\right)^{8} \quad \left(\frac{1}{2}\right)^{10-8} \quad + 10_{c_9} \left(\frac{1}{2}\right)^{9} \quad \left(\frac{1}{2}\right)^{10-9} \quad + 10_{c_{10}} \left(\frac{1}{2}\right)^{10} \quad \left(\frac{1}{2}\right)^{10-10}$$

$$= 10_{c_7} \left(\frac{1}{2}\right)^{7} \quad \left(\frac{1}{2}\right)^{3} + 10_{c_8} \left(\frac{1}{2}\right)^{8} \quad \left(\frac{1}{2}\right)^{2} \quad + 10_{c_9} \left(\frac{1}{2}\right)^{9} \quad \left(\frac{1}{2}\right) \quad + 10_{c_{10}} \left(\frac{1}{2}\right)^{10}$$

$$= \left(\frac{1}{2}\right)^{10} \quad \left(10_{c_7} + 10_{c_8} + 10_{c_9} + 10_{c_{10}}\right)$$

$$= \frac{11}{64} = 0.1719$$

(ii) Probability of getting atleast 6 heads is given by

$$P(X \geq 6) = \text{P(X = 6)} + \text{P(X = 7)} + \text{P(X = 8)} + \text{P(X = 9)} + \text{P(X = 10)}$$

$$=\left(\frac{1}{2}\right)^{6}\ \left(\frac{1}{2}\right)^{10-6}+{}^{10}c_{7}\left(\frac{1}{2}\right)^{7}\ \left(\frac{1}{2}\right)^{10-7}+{}^{10}c_{8}\left(\frac{1}{2}\right)^{8}\ \left(\frac{1}{2}\right)^{10-8}+{}^{10}c_{9}\left(\frac{1}{2}\right)^{9}\ \left(\frac{1}{2}\right)^{10-9}+{}^{10}c_{10}\left(\frac{1}{2}\right)^{10}\ \left(\frac{1}{2}\right)^{10-10}$$

=0.3769

2)If the probability of defective bolt is 0.2

(i)Find mean

(ii)Standard deviation

For the distribution of bolts in a total of 400.

Sol: n=400, p=0.2, q=1-p=1-0.2=0.8

(i)Mean=np

=(400)(0.2)

=80

(ii)Variance=npq

=(400)(0.2)(0.8)

=64

Standard deviation $=\sqrt{npq}$

$=\sqrt{64}$

$=8$

3) Out of 800 families with 5 children each, how many would you expect to have

a) 3 boys

b) 5 girls

c) either 2 or 3 boys

d) atleast one boy

Assume equal probabilities for boys and girls

Sol: Let the number of boys (in) each family is X

Given p=1/2 and q=1/2, n=5

The probability distribution is

P(x)= $p(x)=n_{c_x}p^{x}\ \ q^{n-x}$

$=5_{c_x}\left(\frac{1}{2}\right)^{x}\ \ \left(\frac{1}{2}\right)^{n-x}$

(a)P (3 boys)=P(X=3)

$$= 5c_3 \left(\frac{1}{2}\right)^3 \quad \left(\frac{1}{2}\right)^{5-3}$$

$$= 5c_3 \left(\frac{1}{2}\right)^5$$

$$= \frac{5}{16}$$

Thus for 800 families the probability of number of families having 3 boys

$$\frac{5}{16} \times 800 = 250 \text{ families}$$

3 Boys $\longrightarrow 250$ families

(b) p(5 girls)=P(x=0)=p(no boys)

$$= 5c_0 \left(\frac{1}{2}\right)^0 \quad \left(\frac{1}{2}\right)^{5-0}$$

$$= \left(\frac{1}{2}\right)^5 = \frac{1}{32}$$

Thus for 800 families the probability of no. of families having 5 girls $= \left(\frac{1}{32}\right) \times 800$

=25 families

(c) P(either 2 or 3 boys)=p(x=2)+p(x=3)

$$\Longrightarrow 5c_2 \left(\frac{1}{2}\right)^2 \left(\frac{1}{2}\right)^3 + 5c_3 \left(\frac{1}{2}\right)^3 \quad \left(\frac{1}{2}\right)^2$$

$$= 5c_2 \left(\frac{1}{2}\right)^5 + 5c_3 \left(\frac{1}{2}\right)^5$$

$$= \frac{5}{8} \times 800$$

$\Longrightarrow$500 families

(d) atleast one boy

P($x \geq 1$)=1-P(x=0)

$p = (1 - \frac{1}{2^5}) 800$

=775

4) Mean and variance of the binominal distribution is 4 and 4/3 respectively.

Find P(X ≥ 1)

Sol: Given Mean=np=4 $\longrightarrow(1)$

Variance=npq=4/3 $\longrightarrow(2)$

$(2) \div (1)$

$\frac{npq}{np} = \frac{4/3}{4} = \frac{1}{3}$

q=1/3

P=1-q = 2/3

np=4

n(2/3)=4

n=6

$P(X \geq 1) = 1 - P(X = 0)$

$= 1 - 6_{c_0} (2/3)^0 \quad (1/3)^{6-0}$

$= 1 - (1/3)^6$

$= \frac{728}{729}$

5) Determine the probability of getting the sum exactly 3 times in 7th rows with a pair of fair dice.

Sol: In a single throw of a pair of four dice a sum in 6 can occur in 5 ways i.e (1,5)(5,1)(2,4)(3,3) and (4,2) out of 36 ways.

Thus P=5/36 and q=31/36 (∴p+q=1)

N=7[trails]

∴Probability of getting 6 exactly thrice in 7 throws= $= 7_{c_3} \times p^3 q^{7-3}$

$= 7_{c_3} \left(\frac{5}{36}\right)^3 \left(\frac{31}{36}\right)^4$

=0.05155

6) A coin is biased in a way that a head is twice as likely to occur as a tail. If the coin is tossed 3 times find the probability of getting 2 tail and 1 head.

Sol: Given P(H)=2 P(T)

We know that

P(H)+P(T)=1

2P(T)+P(T)=1

3P(T)=1

P(T)=1/3

P(H)=2/3

Let getting a tail is a success and getting a head is a failure the p=1/3 and q=2/3

Here n=3, x=2

$\therefore$Required Probability$= = 3_{c_2}\left(\frac{1}{3}\right)^2\left(\frac{2}{3}\right)^{3-2}$

$= 3_{c_2}\left(\frac{1}{3}\right)^2\left(\frac{2}{3}\right)^1 = 2/9$

Fitting a Binomial Distribution

(1) Fit binomial distribution to the following frequency distribution

x	0	1	2	3	4	5	6
f	13	25	52	58	32	16	4

Sol: Where n=number of trails=6

$N = \sum f = 200$

$\therefore$Mean$= \frac{\sum f_x}{\sum f} = \frac{0\times13+1\times25+2\times52+3\times58+4\times32+5\times16+6\times4}{200}$

$= \frac{535}{200}$=2.675

$\therefore$np=2.675

6p=2.675

p=0.446

q=1-p=1-0.446=0.554

Fit a Binomial distribution

x	f observed frequency	$P(x) = n_{c_x} p^x \quad q^{n-x}$ $= 6_{c_x} (0.446)^x \quad (0.554)^{6-x}$	Expected frequency f(x)=NP(x) =200 P(x)
0	13	$P(0) = 6_{c_0} (0.446)^0 \quad (0.554)^{6-0} = 0.028$	$5.78 \cong 6$
1	25	$P(1) = 6_{c_1} (0.446)^1 \quad (0.554)^{6-1} = 0.139$	$27.9 \cong 28$
2	52	$P(2) = 6_{c_2} (0.446)^2 \quad (0.554)^{6-2} = 0.281$	$56.21 \cong 56$
3	58	$P(3) = 6_{c_3} (0.446)^3 \quad (0.554)^{6-3} = 0.301$	$60.3 \cong 60$
4	32	$P(4) = 6_{c_4} (0.446)^4 \quad (0.554)^{6-4} = 0.182$	$36.4 \cong 36$
5	16	$P(5) = 6_{c_5} (0.446)^5 \quad (0.554)^{6-5} = 0.058$	$11.74 \cong 12$
6	4	$6_{c_6} (0.446)^6 \quad (0.554)^{6-6} = 0.0079$	$1.57 \cong 2$
	200		200

HW

1) 4 coins are tossed 160 times. The number of times x heads occur is given below.

x	**0**	**1**	**2**	**3**	**4**
f	**8**	**34**	**69**	**43**	**6**

2) The Probability that John hits the target is ½ He fires 6 times. Find the probability that he hits the target

(a)Exactly 2 times

(b)More than 4 times

(c) At least once

1) Sol: p=1/2, q=1//2, n=4, N=160

x	P	$P(x) = n_{c_x} p^x \quad q^{n-x}$	Expected frequency f(x)=NP(x) =160 P(x)
0	8	P(0)= $4_{c_0} p^0 \quad (q)^{4-0} = \left(\frac{1}{2}\right)^4$	10
1	34	P(1)= $4_{c_1} p^1 \quad (q)^{4-1} = \left(\frac{1}{4}\right)$	40
2	69	P(2)= $4_{c_2} p^2 \quad (q)^{4-2} = \left(\frac{3}{8}\right)$	60
3	43	P(3)= $4_{c_3} p^3 \quad (q)^{4-3} = \left(\frac{1}{4}\right)$	40
4	6	P(4)= $4_{c_4} p^4 \quad (q)^{4-4} = \left(\frac{1}{16}\right)$	10
	160		160

2) Sol: Probability of hitting a target=p=1/2

Probability of no hit q=1/2

Number of trails=n=6

Number of hits(Sucesses)=X

(i) P(exactly 2 times)=P(X=2)

$= 6_{c_2} \left(\frac{1}{2}\right)^2 \quad \left(\frac{1}{2}\right)^4$

$= \frac{15}{2^6} = 0.234$

(ii) P(More than 4 times)=P(x>4)

=P(X=5)+P(X=6)

$$= 6c_5 \left(\frac{1}{2}\right)^5 \left(\frac{1}{2}\right) + 6c_6 \left(\frac{1}{2}\right)^6 \left(\frac{1}{2}\right)^0$$

$$= \frac{6}{2^6} + \frac{1}{2^6} = \frac{7}{2^6} = 0.1094$$

(iii)P(atleast once)=P($X \geq 1$)=1-P(X=0)

$$= 1 - 6c_0 \left(\frac{1}{2}\right)^0 \left(\frac{1}{2}\right)^6$$

$$= 1 - \frac{1}{2^6} = 0.9844$$

4.5 POISSON DISTRIBUTION

Definition: A Random variable X is said to follow a Poisson distribution if it assumes only non-negative values and its probability mass function is given by

$$P(x, \lambda) = P(X=x) = \begin{cases} \frac{e^{-\lambda}\lambda^x}{x!} & x = 0,1,2,....\lambda > 0 \\ 0 & otherwise \end{cases}$$

Here $\lambda > 0$ is called the parameter of the distribution.

Note:

$$\sum_{x=0}^{\infty} P(X = x) = \sum_{x=0}^{\infty} \frac{e^{-\lambda}\lambda^x}{x!}$$

$$= e^{-\lambda} \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} \quad \Big\} by \quad e^x \quad exception$$

$$= e^{-\lambda}.e^{\lambda} = 1$$

This is known as probability function.

2) The distribution function is $F(x)=P(X \leq x) = \sum_{r=0}^{x} P(r)$

$$= \sum_{r=0}^{x} \frac{e^{-\lambda}.e^{\lambda}}{r!}$$

$$= e^{-\lambda} \sum_{r=0}^{x} \frac{\lambda^r}{r!}, r = 0,1,2....$$

Mean=Variance

4.6 CONSTANTS OF POISSON DISTRIBUTION

(i)Mean

$$\mu = E(x) = \sum_{x=0}^{\infty} x\ p(\mathrm{x})$$

$$\mu = E(x) = \sum_{x=0}^{\infty} x.\ \frac{e^{-\lambda}\lambda^{x}}{x!}$$

$$= e^{-\lambda} \sum_{x=0}^{\infty} x.\ \frac{\lambda.\lambda^{x-1}}{x(x-1)!}$$

$$= \lambda e^{-\lambda} \sum_{x=1}^{\infty} \frac{\lambda^{x-1}}{(x-1)!}$$

$$= \lambda e^{-\lambda}\left[1+\frac{\lambda}{1!}+\frac{\lambda^2}{2!}+\frac{\lambda^3}{3!}+.....\right]$$

$$= \lambda e^{-\lambda} e^{\lambda}\left[\therefore e^{x} = 1+x+\frac{x^2}{2!}+.....\right]$$

$$= \lambda$$

∴ Mean of Poisson distribution is λ

(ii)Variance

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

Consider $E(\mathrm{X}^2) = \sum_{x=0}^{\infty} x^2 p(x)$

$$= \sum_{x=0}^{\infty} (x(\mathrm{x}-1)+\mathrm{x})\ \frac{e^{-\lambda}\lambda^{x}}{x!}$$

$$= \sum_{x=0}^{\infty} x(\mathrm{x}-1)\ \frac{e^{-\lambda}\lambda^{x}}{x!} + \sum_{x=0}^{\infty} x\ \frac{e^{-\lambda}\lambda^{x}}{x!}$$

$$= \sum_{x=0}^{\infty} x(\mathrm{x}-1)\ \frac{e^{-\lambda}\lambda^{x-2}}{x(x-1)(x-2)!} + \lambda$$

$$= e^{-\lambda}\lambda^{2} \sum_{x=2}^{\infty} \frac{\lambda^{x-2}}{(x-2)!} + \lambda$$

$$= e^{-\lambda}\lambda^2\left[1+\lambda+\frac{\lambda^2}{2!}+.....\right]+\lambda$$

$$= e^{-\lambda}\lambda^2.e^{-\lambda}+\lambda$$

$$= E(X^2) = \lambda^2+\lambda$$

V(x)= $\lambda^2+\lambda-\lambda^2$

$= \lambda$

∴Mean=Variance= λ

Standard deviation of Poisson Distribution= $\sqrt{\lambda}$

(iii) Recurrence relation for Poisson distribution

We have

P(x)= $\frac{e^{-\lambda}\lambda^x}{x!}$

P(x+1)= $\frac{e^{-\lambda}\lambda^x}{x+1)!}$

P(x+1)= $\frac{e^{-\lambda}\lambda^x.\lambda}{(x+1)\,x!}$

$= \frac{e^{-\lambda}\lambda^x}{x!}\,\frac{\lambda}{(x+1)}$

P(x+1)=P(x). $\frac{\lambda}{(x+1)}$

$\Longrightarrow$ P(x+1)= $\frac{\lambda}{(x+1)}$P(x)

(iv)Moment generating function of Poisson distribution

$$M_x(t) = \sum_{x=0}^{\infty} e^{-tx}p(t)$$

$$= \sum_{x=0}^{\infty} e^{-tx}\frac{e^{-\lambda}\lambda^x}{x!}$$

$$= \sum_{x=0}^{\infty} e^{-\lambda} \frac{(e^{t\lambda}.\lambda)^x}{x!}$$

$$= e^{-\lambda}\left[1+\frac{\lambda e^t}{1!}+\frac{(\lambda e^t)^2}{2!}+.....\right]$$

$$= e^{-\lambda} e^{\lambda e^t}$$

$$M_x(t) = e^{\lambda(e^t-1)}$$

4.6 WORKED OUT PROBLEMS

1) If the probability that an individual suffers a bad reaction from a certain injection is 0.001. Determine the probability that out of 2000 individuals

(i) Exactly 3

(ii) More than 2000 individuals

(iii) None

(iv) More than one individual suffered by 1 individual.

Sol: P=0.001, n=2000

$\lambda = np$

=(0.001)(2000)

=2

(i) P(X=x)=$\frac{e^{-\lambda}\lambda^x}{x!}$

$$P(3) = P(\text{x} = 3) = \frac{e^{-2}\lambda^3}{3!}$$

$$= 0.1804$$

(ii) P(more than 2)= $P(\text{x} \geq 2)$

$$=1\text{-}[\text{P(x=0)+P(x=1)}]\text{-P(x=2)}]$$

$$=1-\left[\frac{e^{-2}2^0}{0!}+\frac{e^{-2}2^1}{1!}+\frac{e^{-2}2^2}{2!}\right]$$

$$=1-e^{-2}[1+2+2]$$

$$=1-e^{-2}[5]$$

$=0.3233$

(iii)P(none)=P(x=0)= $\frac{e^{-2}2^0}{0!}$

$=0.1353$

(iv) P(more than one)=P(x>1)

$=1-[P(x=0)+P(x=1)]$

$=1-\left[\frac{e^{-2}2^0}{0!}+\frac{e^{-2}2^1}{1!}\right]$

$=1-e^{-2}[1+2]$

$==1-3e^{-2}$

$=0.594$

2) If a random variable has Poisson distribution such that P(1)=P(2) find (i) mean of the distribution

(ii)P(4) (iii)P(x ≥ 1) (iv)P(1<x<4)

Sol: Given P(1)=P(2)

$$\frac{e^{-\lambda}\lambda^1}{1!}=\frac{e^{-\lambda}\lambda^2}{2!}$$

$$1=\frac{\lambda}{2!}$$

$$\lambda=2$$

(i)mean=2

(ii) $\frac{e^{-2}2^4}{4!}$ =0.0902

(iii)P(x ≥ 1)=1-P(x<1)

=1-P(x=0)

=1- $\frac{e^{-2}2^0}{0!}$

$=1-e^{-2}$

=0.8647

(iv)P(1<x<4)=P(x=2)+P(x=3)

$$=\frac{e^{-2}-2^2}{2!}+\frac{e^{-2}-2^3}{3!}$$

=0.4511

3) Using recurrence formula find the probabilities when x=0,1,2,3,4, and 5. If the mean of Poisson distribution is 3.

Sol: Given mean of the Poisson distribution is 3 i.e λ =3

Now the Poisson distribution is P(x)= $\frac{e^{-\lambda}\lambda^x}{x!}$

$$=\frac{e^{-3}-3^x}{\text{x!}}$$ (can't apply recurrence formula ∴ /0)

$\therefore$P(x=0)=$\frac{e^{-3}-3^0}{0!}=e^{-3}$=0.498

Using recurrence formula

P(x+1)=$\frac{\lambda}{x+1}p(x)$ $\longrightarrow(1)$

Put x=0 in (1)

P(1)=$\frac{3}{1}p(0)$=3(0.0498)=0.1494

Put x=1 in (1), we have

P(2)=$\frac{3}{1+1}p(1)$

$=\frac{3}{2}(0.1494)$

=0.2241

Put x=2 in (1), we have

P(3)=$\frac{3}{2+1}$p(2)

=0.2241

Put x=3 in (1),

$$P(4)=\frac{3}{3+1}p(3)$$

$$=\frac{3}{4}(0.2241)$$

=0.1681

Put x=4 in (1)

$$P(5)=\frac{3}{4+1}p(4)$$

$$=\frac{3}{5}(0.1681)$$

=0.1009

4) Suppose 2% of the people on the average are left handed. Find (i) the probability of finding 3 are more left handed.

(ii)The probability of finding none are 1 left handed

Sol: Let x be the number of left handed.

Given mean= λ =2%=0.02

We have $P(x)=\frac{e^{-\lambda}\lambda^{x}}{x!}$

$$=\frac{e^{-0.02}-(0.02)^{x}}{x!}$$

(i) $P(x\geq 3)$ =1-P(x<3)

=1-[P(x=0)+P(x=1)+P(x=2)]

$$=1-\left[\frac{e^{-0.02}-(0.02)^{0}}{0!}+\frac{e^{-0.02}-(0.02)^{1}}{1!}+\frac{e^{-0.02}-(0.02)^{2}}{2!}\right]$$

$$=1.3077\times10^{-6}$$

(ii) $p(x\leq 1)=p(0)+p(1)$

$$=\frac{e^{-0.02}-(0.02)^0}{0!}+\frac{e^{-0.02}-(0.02)^1}{1!}$$

=0.999

5)If x is a Poisson variate such that 3p(x=4)=1/2 p(x=2)+p(x=0)

Find (i)the mean (ii)p(x ≤ 2)

Sol: Given

$$3p(x=4)=\frac{1}{2}\mathrm{p(x=2)+p(x=0)}$$

$$\frac{3e^{-\lambda}\lambda^4}{4!}=\frac{1}{2}\frac{e^{-\lambda}\lambda^2}{2!}+\frac{e^{-\lambda}\lambda^0}{0!}$$

$$e^{-\lambda}\left[\frac{3\lambda^4}{24}\right]=e^{-\lambda}\left[\frac{\lambda^2}{4}+1\right]$$

$$\Longrightarrow\frac{\lambda^2}{8}=\frac{\lambda^2}{4}+1$$

$$=\frac{\lambda^4}{8}-\frac{\lambda^2}{4}-1=0$$

$$=\frac{\lambda^4-2\lambda^2-8}{8}=0$$

$$\Longrightarrow\lambda^4-2\lambda^2-8=0$$

$$\lambda^4-4\lambda^2+2\lambda^2-8=0$$

$$\lambda^2(\lambda^4-4)+2(\lambda^4-4)=0$$

$$(\lambda^2+2)(\lambda^4-4)=0$$

$$\lambda=\pm 2$$

$$\lambda=2\quad(\because\lambda\geq 0)$$

(i)mean= $\lambda=2$

$$\mathrm{(ii)\ p(x\leq 2)=p(x=0)+p(x=1)+p(x=2)}$$

$$=\frac{e^{-2}2^0}{0!}+\frac{e^{-2}2^1}{1!}+\frac{e^{-3}2^2}{2!}=0.66$$

6) If 2% of the light bulbs are defective find

(i)at least 1 defective

(ii) Exactly 7 are defective

(iii)P(1<x<8) in a sample of 100

Sol: Given P=2%=0.02

n=100

mean= λ =np=0.02×100=2

(i)at least 1 defective

$$P(X\geq 1)=1-P(X\geq 0)$$

$$=1-\left[\frac{e^{-2}1}{1}\right]$$

$$=1-\left[e^{-2}\right]$$

=0.864764

(ii) Exactly 7 are defective

$$P(\mathrm{x}=7)=\frac{e^{-2}7}{7!}$$

=0.00001

(iii) P(1<x<8)

=P(x=2)+P(x=3)+P(x=4)+ P(x=5)+P(x=6)+P(x=7)

$$=\frac{e^{-2}.2}{2!}+\frac{e^{-2}.3}{3!}+\frac{e^{-2}.4}{4!}+\frac{e^{-2}.5}{5!}+\frac{e^{-2}.6}{6!}+\frac{e^{-2}.7}{7!}=0.593$$

Fit a Poisson Distribution :

1) 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

Sol: n=4

$N = \sum f = 200$

$$\therefore \lambda = \text{Mean} = \frac{\sum f_x}{\sum f} = \frac{0\times109 + 1\times65 + 2\times22 + 3\times3 + 4\times1}{200}$$

$$= \frac{65+44+9+4}{200}$$

$$= 0.61$$

$$\therefore \lambda = 0.61$$

Fit a Poisson distribution

x	Observed frequency f	$P(x) = \frac{e^{-\lambda}\lambda^x}{x!} = \frac{e^{-0.61}.(0.61)^x}{x!}$	Expected frequency F(x)=N.P(x) =200 P(x)
0	109	$P(0) = \frac{e^{-0.61}.(0.61)^0}{0!} = 0.543$	$0.543\times200 = 109$
1	65	$P(1) = \frac{e^{-0.61}.(0.61)^1}{1!} = 0.331$	$66.2 \cong 66$
2	22	$P(2) = \frac{e^{-0.61}.(0.61)^2}{2!} = 0.1011$	$20.2 \cong 20$
3	3	$P(3) = \frac{e^{-0.61}.(0.61)^3}{3!} = 0.0205$	$4.1 \cong 4$
4	1	$P(4) = \frac{e^{-0.61}.(0.61)^4}{4!} = 0.0031$	$0.62 \cong 1/2$

2) Fit a Poisson distribution to the following

x	0	1	2	3	4	5
f	142	156	69	27	5	1

Sol: n=5

$N = \sum f = 400$

$$\therefore \lambda = \text{Mean} = \frac{\sum fx}{\sum f} = \frac{0\times142 + 1\times156 + 2\times69 + 3\times27 + 4\times5 + 5\times1}{400}$$

$$= \frac{400}{400}$$

$$= 1$$

$$\therefore \lambda = 1$$

Fit a Poisson distribution

x	Observed frequency f	$P(x) = \frac{e^{-\lambda}\lambda^x}{x!}$ $P(x) = \frac{e^{-1}1^x}{x!}$	Expected frequency F(x)=N.P(x) =400 P(x)
0	142		147
1	156		147
2	69		74
3	27		25
4	5		6
5	1		1

4.8 SUMMARY

This chapter explores discrete probability distributions, focusing on two fundamental types: binomial and Poisson distributions. The binomial distribution models the probability of a fixed number of successes in a series of independent trials, with key constants such as mean, variance, and standard deviation. Several worked-out problems illustrate its practical applications. The Poisson distribution is introduced as a limiting case of the binomial distribution, used to model rare events occurring within a fixed interval of time or space. The chapter also discusses its constraints and includes solved examples to strengthen understanding. These concepts are widely used in real-world applications like risk assessment, reliability engineering, and queuing theory.

4.9 TECHNICAL TERMS

- Discrete Probability Distribution
- Binomial Distribution
- Bernoulli Trials
- Mean and Variance of Binomial Distribution
- Poisson Distribution
- Lambda (λ) – Poisson Parameter
- Probability Mass Function (PMF)
- Moment-Generating Function (MGF)

4.10 SELF-ASSESSMENT QUESTIONS

Short Questions

1. What is a discrete probability distribution? Provide an example.
2. Define a binomial distribution and state its probability mass function (PMF).
3. What are the key constants (mean and variance) of a binomial distribution?
4. Define the Poisson distribution and mention its key parameter.
5. How is the Poisson distribution derived from the binomial distribution?

Essay Questions

1. Explain the binomial distribution with its assumptions, formula, and real-life applications.
2. Derive the mean and variance of the binomial distribution.
3. Describe the Poisson distribution, its assumptions, and its significance in modeling rare events.
4. Compare and contrast binomial and Poisson distributions with examples.
5. Solve a real-world problem using both binomial and Poisson distributions, explaining when each is appropriate.

4.11 FURTHER READINGS

1. "An Introduction to Probability Theory and Its Applications, Volume 1" – William Feller, Wiley, 1968.
2. "Probability Theory: A Concise Course" – Y. A. Rozanov, Dover Publications, 1977.
3. "Probability Theory: An Introductory Course" – Iakov G. Sinai, Springer, 1992.
4. "Probability Theory: A Comprehensive Course" – Achim Klenke, Springer, 2020.
5. "Fat Chance: Probability from 0 to 1" – Benedict Gross, Joe Harris, Emily Riehl, Cambridge University Press, 2019.

LESSON-5

CONTINUOUS PROBABILITY DISTRIBUTION

OBJECTIVES:

After going through this lesson, you will be able to

- Understand Discrete Probability Distributions – Learn the concept of discrete probability distributions and their role in statistical modelling.
- Explore Binomial Distribution – Study the properties, probability mass function (PMF), and real-world applications of binomial distribution.
- Analyze Constants of Binomial Distribution – Understand key parameters such as mean, variance, and standard deviation and their significance.
- Apply Binomial Distribution Through Worked-Out Problems – Solve numerical problems to gain practical understanding of the binomial model.
- Learn Poisson Distribution – Understand the Poisson distribution as a limiting case of the binomial distribution and its use in modeling rare events.
- Study Constraints of Poisson Distribution – Identify the conditions under which the Poisson distribution is applicable.
- Practice Problem-Solving – Work on solved examples to enhance problem-solving skills related to binomial and Poisson distributions.

STRUCTURE OF THE LESSION:

5.1 Continuous Probability Distribution
5.2 Properties of Probability density function
5.3 Probability distribution function
5.4 Parameters
5.5 Worked out Problems
5.6 Some Theorems
5.7 Exponential Distribution
5.8 Summary
5.9 Technical Terms
5.10 Self-Assessment Questions
5.11 Further Readings

5.1 CONTINUOUS PROBABILITY

When a random variable X Takes every value in an interval to gives rise to continuous distribution of x. The distribution defined by variates like temperature, heights and weights are continuous distributions.

5.2 PROBABILITY DENSITY FUNCTION

For continuous variable the probability distribution is called probability density function because it is defined for every point in the range and not only for certain values and it is denoted by f(x).

Properties:

1) $f(x)\geq 0$; $-\infty<x<\infty$

2) $\int_{-\infty}^{\infty} f(x)dx = 1$ ($\therefore\infty$=Substitute given limits)

3)$P(a \leq x \leq b) = \int_a^b f(x)dx$

5.3 PROBABILITY DISTRIBUTION FUNCTION

Distribution function of a continuous random variable x is denoted by F(x) and is defined as

$F(x)=P(x \leq x) = \int_{-\infty}^{x} f(x)dx$

Properties:

1)$0\leq F(x) \leq 1$, $-\infty<x<\infty$

2)$F^1(x)=f(x)$

3)$F(-\infty)=0$

4)$F(\infty)=1$

5.4 MEASURES OF CENTRAL TENDENCY FOR CONTINUOUS PROBABILITY DISTRIBUTION:

(i) Mean:

Mean of the distribution is given by

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

If x is defined by A to B then

$$\mu = E(x) = \int_a^b xf(x)dx$$

In general

Mean or expection of any function $\boldsymbol{\phi}(x)$ is given by

$$E(\boldsymbol{\phi}(x)) = \int_{-\infty}^{\infty} \phi(x)f(x)dx$$

(ii) Median:

Median is the point which ÷ the entire distribution into 2 equal parts.

In case of continuous distribution median is the point which divides the total area into 2 equal parts.

∴ Thus if x is defined from A to B and M is the median then

$$\int_a^M f(x)dx = \int_a^b f(x)dx = 1/2$$

Solving M we get the median.

(iii) Mode:

Mode is the value of x for which f(x) is maximum.

Mode is thus given by

$f^1(x)=0$ and $f^{11}(x)<0$ for $a<x<b$

(iv) Variance:

$$\sigma^2 = v(x) = \int_{-\infty}^{\infty}(x-\mu)^2 \, f(x)dx$$

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

$$= E(x^2) - (E(x))^2$$

(v) Mean Deviation:

Mean deviation about the mean μ is given by

$$\int_{-\infty}^{\infty} |x-\mu| f(x)dx$$

5.5 WORKED OUT PROBLEMS

1)If a probability density of a random variable is given by $f(1-x^3)$

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

find the value of k and the probabilities that a random variable having this probability density will take by value (i)Between 0.1 and 0.2

(ii)>0.5

Solution: Given

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

We have $\int_{-\infty}^{\infty} f(x)dx$

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

$0+\int_0^1 k(1-x^2)dx+0 = 1$

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

$k[(1-1/3)-(0-0)] = 1$

$k - 2/3 = 1$

$k = 3/2$

$\therefore f(x) = \begin{cases} 3/2\left(1-x^2\right) & \text{for } 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$

(i) The probability that the variate will take on a value between 0.1 and 0.2 is $(1-x^2)$

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

$= \int_{0.1}^{0.2} \frac{3}{2}(1-x^2)dx$

$= \frac{3}{2}\left[x - \frac{x^3}{3}\right]_{0.1}^{0.2}$

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

$=0.1465$

(iii) $P(x>0.5) = \int_{0.5}^{\infty} f(x)dx$

$= \int_{0.5}^{0.1} f(x)dx + \int_{1}^{\infty} f(x)dx$

$= \int_{0.5}^{1} \frac{3}{2}(1-x^2)dx + 0$

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

$=0.312$

2)The probability density function f(x) of a continuous random variable is given by

$f(x) = ce^{-|x|}$, $-\infty<x<\infty$ show that $c=\frac{1}{2}$ and find the mean and variance of the distributions. Also find the probability that the variate lies between 0 and 4.

Solution: Given f(x)= $c\,e^{-|x|}$ [∴ 'e' even functions]

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

$\Rightarrow \int_{-\infty}^{\infty} c\,e^{-|x|}\,dx = 1$

$\Rightarrow 2c\int_{0}^{\infty} e^{-|x|}\,dx = 1$ [∴ Since $e^{-|x|}$ is even]

$\Rightarrow 2c\int_{0}^{\infty} e^{-x}\,dx = 1$ [∴ |x|=x]

$\Rightarrow 2c\left[\frac{e^{-x}}{-1}\right]_0^{\infty} = 1$

$\Rightarrow$ 2c[0-1]=1

$\Rightarrow$ 2c=1

$\Rightarrow$ c=1/2

(ii)Mean= $\mu = E(x) = \int_{-\infty}^{\infty} x\,f(x)dx$

$= \int_{-\infty}^{\infty} x\,e^{-|x|}\,dx$ [$\int_{-a}^{a} f(x)dx = 0$ if f(x) is odd $= 2\int_{-a}^{a} f(x)dx$ is $f(x)$is even]

=0[Integrand to odd]

(iii)Variance

$\sigma^2 = \int_{-\infty}^{\infty}(x-\mu)^2\,f(x)dx$

$=\int_{-\infty}^{\infty}(x-0)^2 .. \frac{1}{2}e^{-|x|}\,dx$

$=\int_{-\infty}^{\infty}x^2 . e^{-|x|}\,dx$

$=\frac{1}{2}\int_{-\infty}^{\infty}x^2 . e^{-|x|}\,dx$

$=2.\frac{1}{2}\int_{0}^{\infty}x^2 . e^{-|x|}\,dx$, Since integrand is even

$=\int_{0}^{\infty}x^2 e^{-x}dx = \left[x^2.\frac{e^{-x}}{-1} - 2x\frac{e^{-x}}{-1} + 2\frac{e^{-x}}{-1}\right]_0^{\infty}$

=[0-(-2)]=2

(iii)The probability between 0 and 4 is =P(0≤x≤4)

$=\frac{1}{2}\int_{0}^{4} e^{-|x|}\,dx$

$= \frac{1}{2}\int_0^4 e^{-x}\ dx$ [∴ in 0<x<4, |x|=x]

$= -\frac{1}{2}\left(e^{-x}\right)_0^4$

$= -\frac{1}{2}\left(e^{-4}-1\right)$

$= \frac{1}{2}\left(1-e^{-4}\right)$

$=0.4908$

3) A continuous random variable has probability density function

$$f(x)=\begin{cases} kx\ e^{-\lambda x} & \text{for } x\geq 0,\ \lambda > 0 \\ 0 & \text{otherwise} \end{cases}$$

λ is parameter

Determine (i)k (ii)Mean (iii)Variance

Solution: We know $\int_{-\infty}^{\infty} x\, f(x)dx=1$ [∴ Since the total probability is unity]

i.e $\int_{-\infty}^{0} 0.\,dx+\int_0^{\infty} k\,x\ e^{-\lambda x}\,dx=1$ i.e $k\int_0^{\infty} x\ e^{-\lambda x}\,dx=1$

i.e $k\left[x\left(\frac{e^{-\lambda x}}{-\lambda}\right)-1\left(\frac{e^{-\lambda x}}{-\lambda^2}\right)\right]_0^{\infty}=1$

i.e $k\left[\left(0-0\right)-1\left(0-\frac{1}{\lambda^2}\right)\right]=1$ or $k=\lambda^2$

now f(x) becomes $f(x)=\begin{cases} \lambda^2 x e^{-\lambda x} & \text{for } x\geq 0,\ \lambda > 0 \\ 0, & \text{otherwise} \end{cases}$

(ii) Mean of the distribution, $\mu=\int_{-\infty}^{\infty} x\, f(x)dx$

i.e $\int_{-\infty}^{0} 0.\,dx+\int_0^{\infty} x\,.\,\lambda^2\ x e^{-\lambda x}\,dx=\lambda^2\int_0^{\infty} x^2\ e^{-\lambda x}\,dx$

$=\lambda^2\left[x^2\left(\frac{e^{-\lambda x}}{-\lambda}\right)-2x\left(\frac{e^{-\lambda x}}{\lambda^2}\right)+2\left(\frac{e^{-\lambda x}}{-\lambda^3}\right)\right]_0^{\infty}$

$=\lambda^2\,[(0-0+0)-(0-0-2/\lambda^3)=\frac{2}{\lambda}$

(iii) Variance of the distribution

$$\sigma^2 = \int_0^\infty x^2 \, f(x)dx - \mu^2$$

i.e $\sigma^2 = \int_0^\infty x^2 \, f(x)dx - \left(\frac{2}{\lambda}\right)^2 = \lambda^2 \int_0^\infty x^3 \; e^{-\lambda x} \, dx - 4/\lambda^2$

$$= \lambda^2 \left[x^3 \left(\frac{e^{-\lambda x}}{-\lambda x} \right) - 3x^2 \left(\frac{e^{-\lambda x}}{\lambda^2} \right) + 6x \left(\frac{e^{-\lambda x}}{-\lambda^3} \right) - 6 \left(\frac{e^{-\lambda x}}{-\lambda^4} \right) \right]_0^\infty - \frac{4}{\lambda^2}$$

$$= \lambda^2 \, [(0\text{-}0\text{+}0\text{-}0)\text{-}(0\text{-}0\text{+}0\text{-}6/\lambda^4)] - \frac{4}{\lambda^2}$$

$$= \frac{6}{\lambda^2} - \frac{4}{\lambda^2} = \frac{2}{\lambda^2}$$

4) A continuous random variable x has the distribution function

$$F(x) = \begin{cases} 0 & \text{if } x \le 1 \\ k(x-1)^4 & \text{if } 1 < x \le 3 \\ 1 & \text{if } x > 3 \end{cases}$$

Determine (i) f(x) (ii)k (iii)Mean

Solution: (i) We know that

$$f(x) = \frac{d}{dx} F(x)$$

$$f(x) = \begin{cases} 0 & \text{if } x \le 1 \\ 4k(x-1)^3 & \text{if } 1 < x \le 3 \\ 0 & \text{if } x > 3 \end{cases}$$

(ii) We knew that

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

$$\Rightarrow \int_{-\infty}^{1} f(x)dx + \int_1^3 f(x)dx + \int_3^\infty f(x)dx$$

$$\Rightarrow 0 + \int_1^3 4k(x-1)^3 \, dx + 0 = 1$$

$$\Rightarrow 4k \left[\frac{(x-1)^4}{4} \right]_1^3 = 1$$

$$\Rightarrow k[(3-1)^4 - (3-1)^4] = 1$$

$$\Rightarrow k(16) = 1$$

$$k = 1/16$$

hence

$$f(x)=\begin{cases} 0 & \text{if } x \le 1 \\ \frac{1}{4}(x-1)^3 & \text{if } 1<x\le 3 \\ 0 & \text{Otherwise} \end{cases}$$

(iii)Mean= $\mu=\int_{-\infty}^{\infty} x\, f(x)dx$

$$=\int_{-\infty}^{1} x\, f(x)dx+\int_{1}^{3} x\, f(x)dx+\int_{3}^{\infty} f(x)dx$$

$$=0+\int_{1}^{3} \frac{1}{4}x(x-1)^3\, dx$$

Let x-1=t

dx=dt

Upper limit $\Rightarrow$x=3

$\Rightarrow$t=2

Lower limit $\Rightarrow$x=1

$\Rightarrow$t=0

$$\Rightarrow 1/4\int_0^2 (t+1)t^3 dt$$

$$\Rightarrow 1/4\int_0^2 \left(t^4+t^3\right)dt$$

$$\Rightarrow 1/4\left[\frac{t^5}{5}+\frac{t^4}{4}\right]_0^2$$

$$\Rightarrow 1/4\left[\frac{32}{5}+\frac{16}{4}\right]$$

$$\Rightarrow 1/4\left[\frac{128+80}{20}\right]$$

$$\Rightarrow 1/4\left[\frac{208}{20}\right]$$

$$\Rightarrow \left[\frac{52}{20}\right]$$

$$\Rightarrow \left[\frac{10.4}{4}\right]$$

$\Rightarrow$ 2.6

$\therefore$Mean=2.6

5.6 SOME THEOREMS

1)If x is a continuous random variable and Y is equal to ax+b and prove that E(y)= ae^x+b and variance of Y is equal to a^2(variance of x) i.e a^2v(x) and a,b are constants.

Proof :

Given Y=ax+b $\longrightarrow$(1)

E(Y)=E(ax+b)

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

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

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

E(Y)=aE(x)+b $\longrightarrow$(2)

(1) - (2) gives

Y-E(Y)=(ax+b)-[aE(x)+b]

Y-E(Y)=ax+b-aE(x)-b

Y-E(Y)=a[x-E(x)]

Squaring on both sides

$[Y-E(Y)]^2=a^2[X-E(X)]^2$

Taking expectation on both sides

$E\,[Y-E(Y)]^2=a^2E\,[X-E(X)]^2$

$v(Y)=a^2v(x)$

2)If x is a continuous random variable and k is a constant then prove that

(i)variance(x+k)=v(x)

(ii)$v(kx)=k^2.V(x)$

Solution:

$Var(x)= \int_{-\infty}^{\infty}x^2\, f(x)dx-\int_{-\infty}^{\infty}x\, f(x)dx$

(i) $v(x+k)= \int_{-\infty}^{\infty}(x+k)^2\, f(x)dx-\int_{-\infty}^{\infty}(x+k)\, f(x)dx$

$\int_{-\infty}^{\infty}(x^2+2kx+k^2)\, f(x)dx-[\int_{-\infty}^{\infty}xf(x)dx+ k\int_{-\infty}^{\infty} f(x)dx]$

$\int_{-\infty}^{\infty}x^2\, f(x)dx+2k\int_{-\infty}^{\infty}x\, f(x)dx+k^2\int_{-\infty}^{\infty} f(x)\,dx - [\int_{-\infty}^{\infty}x\, f(x)dx-k]^2$

$E(x^2)+2kE(x)+k^2-[E(x)+k]^2$

$v(x+k)=E(x^2)+2kE(x)+k^2-[E(x)]^2-2KE(x)-k^2$

$=E(x^2)-(E(x))^2$

$v(x+k)=v(x)$.

(ii)$var(kx)= \int_{-\infty}^{\infty} k^2 x^2 \, f(x)dx - [\int_{-\infty}^{\infty} kx \, f(x)dx-k]^2$

$=k^2 \int_{-\infty}^{\infty} x^2 \, f(x)dx-- k^2 [\int_{-\infty}^{\infty} x \, f(x)dx]^2$

$=k^2 E(x^2)-k^2(E(x))^2$

$=k^2 [E(x^2)-(E(x))^2)$

$var(kx)= k^2 v(x)$

3)For the continuous probability function $f(x)=k.x^2.e^{-x}$ when $x\geq 0$ find (i)k (ii)mean (iii)variance

.**Solution:**

(i)We have $\int_{-\infty}^{\infty} f(x)\,dx \quad \therefore \int_0^{\infty} f(x)k.x^2.e^{-x}\,dx=1 \quad (\therefore x\geq 0)$

i.e $k\left[x^2(-e^{-x})-2x(e^{-x)}+2\left(-e^{-x}\right)\right]_0^{\infty}=1$

$k\left[-e^{-x}(x^2+2x+2)\right]_0^{\infty}=1$

$k[0+2]=1$

$k=\frac{1}{2}$

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

$=\int_0^{\infty} k\, x^3 e^{-x}\,dx$

$= k\left[x^3(-e^{-x})-3x^2(e^{-x})+6x\left(-e^{-x}\right)-6e^{-x}\right]_0^{\infty}$

$=k\left[(-e^{-x})[x^3+3x^2+6x+6\right]_0^{\infty}$

$=k[0+6]=6k$

$\therefore \mu=6\left(\frac{1}{2}\right)=3 \qquad (\therefore k=\frac{1}{2})$

(iii)Variance$= \int_{-\infty}^{\infty} x^2 \, f(x)dx-\mu^2$

$=\int_0^{\infty} x^2.\, k\, x^2\, e^{-x}dx-(3)^2$

$=k\int_0^{\infty} x^4\, e^{-x}dx - 9$

$$= k\left[x^4(-e^{-x}) - 4x^3(e^{-x}) + 12x^2\left(-e^{-x}\right) - 24x(-e^{-x}) + 24e^{-x}\right]_0^{\infty} - 9$$

$$= \frac{1}{2}\left[-e^{-x}(x^4 + 4x^3 + 12x^2 + 24x + 24\right]_0^{\infty} - 9$$

$$= \frac{1}{2}[0+24]$$

$$=12-9$$

$$=3$$

5.7 EXPONENTIAL DISTRIBUTION

Definition: A random variable X is said to have an exponential distribution with parameter θ>0; if its p.d.f is given by : $f(x,\theta) = \begin{cases} \theta\, e^{-\theta x}, & x \geq 0 \\ 0, & \text{otherwise} \end{cases}$

The cumulative distribution function F(x) is given by

$$F(x) = f(u)du = \theta \int_0^x e^{(-\theta u)}\, du$$

$$f(x,\theta) = \begin{cases} 1 - \exp(-\theta x), & x \geq 0 \\ 0, & \text{otherwise} \end{cases}$$

Moment generating function of exponential distribution

$$M_x(t) = E(e^{tx}) = \theta \int_0^{\infty} e^{tx}\, e^{-\theta x}\, dx$$

$$= \theta \int_0^{\infty} \exp((-(\theta - t)x)\ dx$$

$$= \frac{\theta}{\theta - t} = \left(1 - \frac{t}{\theta}\right)^{-1} = \sum_{r=0}^{\infty} \left(\frac{t}{\theta}\right)^r, \theta > t$$

$$\therefore \mu_r^{1} = E(x^r) = \text{Coefficient of } \frac{t^r}{r!} \text{ in } M_x(t) = \frac{r!}{\theta^r}, \quad r = 1,2,\ldots$$

Mean $= \mu_1^{1} = \frac{1}{\theta}$ and variance $= u_2$

$$= u_2^{1} - u_1^{1}$$

$$= \frac{2}{\theta^2} - \frac{1}{\theta^2} = \frac{1}{\theta^2}$$

Hence if $X \sim \exp(\theta)$,

Then Mean $= \frac{1}{\theta}$ and Variance $= \frac{1}{\theta^2}$

Remark: Variance$=\frac{1}{\theta^2}=\frac{1}{\theta}.\frac{1}{\theta}=\frac{Mean}{\theta}$

Variance > Mean, if $0<\theta<1$

Variance = Mean, $\theta=1$

Variance < Mean, if $\theta>1$

Hence for the exponential distribution

Variance >, = , or < Mean , for different values of the parameter.

5.8 SUMMARY

This chapter explores continuous probability distributions, which describe random variables that can take any value within a given range. It begins with the probability density function (PDF) and its essential properties, explaining how probabilities are assigned over continuous intervals. The probability distribution function (CDF) is introduced to show cumulative probabilities. Important parameters, such as mean and variance, are discussed to characterize these distributions. Several theorems provide a mathematical foundation for solving probability-related problems. The chapter also covers the exponential distribution, commonly used to model waiting times and failure rates. To reinforce these concepts, worked-out problems illustrate practical applications in real-world scenarios.

5.9 TECHNICAL TERMS

- Continuous Probability Distribution
- Probability Density Function (PDF)
- Cumulative Distribution Function (CDF)
- Mean (Expectation) and Variance
- Normal Distribution
- Exponential Distribution
- Properties of Probability Distributions
- Moment-Generating Function (MGF)

5.10 SELF-ASSESSMENT QUESTIONS

SHORT:

1. What is a continuous probability distribution? Give an example.
2. Define the probability density function (PDF) and state its properties.
3. What is the difference between PDF and cumulative distribution function (CDF)?
4. What are the key parameters of a continuous probability distribution?
5. Define the exponential distribution and mention one of its applications.

ESSAY:

1. Explain the probability density function (PDF) and cumulative distribution function (CDF) with examples.
2. Discuss the properties of a probability density function and its significance.
3. Describe key parameters (mean, variance) of continuous probability distributions and their importance.
4. Explain the exponential distribution, derive its mean and variance, and discuss its applications.
5. Solve a real-world problem involving continuous probability distributions and interpret the results.

5.11 FURTHER READINGS

1. "An Introduction to Probability Theory and Its Applications, Volume 1" – William Feller, Wiley, 1968.
2. "Probability Theory: A Concise Course" – Y. A. Rozanov, Dover Publications, 1977.
3. "Probability Theory: An Introductory Course" – Iakov G. Sinai, Springer, 1992.
4. "Probability Theory: A Comprehensive Course" – Achim Klenke, Springer, 2020.
5. "Fat Chance: Probability from 0 to 1" – Benedict Gross, Joe Harris, Emily Riehl, Cambridge University Press, 2019.

Dr. A.J.V. Radhika

LESSON-6

NORMAL DISTRIBUTION

OBJECTIVES:

After going through this lesson, you will be able to

- Understand the Concept of Normal Distribution: Grasp the fundamental idea behind the normal distribution and its role in statistics.
- Learn the Definition and Key Characteristics: Define the normal distribution and explain its bell-shaped curve, symmetry, and central tendency.
- Analyze the Constants of Normal Distribution: Understand and interpret the key parameters, such as mean and variance, that describe the distribution.
- Examine the Properties of Normal Distribution: Study important properties, including the 68-95-99.7 rule, and how these properties facilitate statistical analysis.
- Apply Knowledge Through Worked-Out Examples: Solve practical problems and examples to reinforce understanding and application of the normal distribution in real-world scenarios.

STRUCTURE OF THE LESSION:

6.1 Normal Distribution
6.2 Constants of normal distribution
6.3 Properties of normal distribution
6.4 Worked out Examples
6.5 Summary
6.6 Technical Terms
6.7 Self-Assessment Questions
6.8 Further Readings

6.1 NORMAL DISTRIBUTION

Now we shall consider continuous distribution, namely the Normal Distribution. A continuous distribution is a distribution in which the variate can take all values within a given range. Examples of continuous distribution are the heights of persons, the speed of a vehicle, etc.

The normal distribution was first discovered by English mathematician De-Moiere (1667-1745) in 1733 and it is also known as Gaussian Distribution. It is another limiting form of the Binomial Distribution for large values of n when neither p nor q is very small and it is derived from the binomial distribution by increasing the number of trails indefinitely.

Definition:

A random variable x is said to have a normal distribution if its density function or probability distribution is given by

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

$$-\infty < x < \infty$$
$$-\infty < \mu < \infty \qquad \sigma > 0$$

Where μ is mean, σ is standard deviation are parameters of the normal distribution.

6.2 CONSTANTS OF NORMAL DISTRIBUTION

(1) Mean: Consider the normal distribution with b, σ as the parameters then

$$f(x:b,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-1/2\left(\frac{x-b}{\sigma}\right)^2}$$

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

$$= \int_{-\infty}^{\infty} x. \frac{1}{\sigma\sqrt{2\pi}} e^{-1/2\left(\frac{x-b}{\sigma}\right)^2} dx$$

Let $Z = \frac{x-b}{\sigma}$

$$x = b + \sigma z$$

$$dx = \sigma dz$$

$$\mu = = \int_{-\infty}^{\infty} (\sigma z)^2 \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-z2}{2}} \sigma\, dz$$

$$= \frac{1}{\sqrt{2\pi}} \left[\int_{-\infty}^{\infty} be^{\frac{-z^2}{2}} dz + \int_{-\infty}^{\infty} \sigma z e^{\frac{-z^2}{2}} dz \right]$$

$$= \frac{1}{\sqrt{2\pi}} \left[b\int_{-\infty}^{\infty} e^{\frac{-z^2}{2}} dz + \sigma \int_{-\infty}^{\infty} z e^{\frac{-z^2}{2}} dz \right]$$

$$= \frac{1}{\sqrt{2\pi}} \left[2b\int_{0}^{\infty} e^{\frac{-z^2}{2}} dz + 0 \right] \left[\begin{array}{l} \therefore e^{\frac{-z^2}{2}} is\ even \\ z e^{\frac{-z^2}{2}} is\ odd \end{array} \right]$$

$$= \frac{2b}{\sqrt{2\pi}} \left[\int_{0}^{\infty} e^{\frac{-z^2}{2}} dz \right]$$

$$= \frac{2b}{\sqrt{2\pi}} \sqrt{\frac{\pi}{2}} \quad \left[\int_{-\infty}^{\infty} e^{\frac{-x^2}{2}} dx = \sqrt{\frac{\pi}{2}} \right]$$

μ =b

(2) Variance of normal distribution

$$V(x)=E(x-b)^2$$

$$= \int_{-\infty}^{\infty} (x-b)^2 f(x)\, dx$$

$$= \int_{-\infty}^{\infty} (x-b)^2 \frac{1}{\sigma\sqrt{2\pi}} e^{-1/2\left(\frac{x-b}{\sigma}\right)^2} \text{ dx}$$

$$\text{Let } Z = \frac{x-b}{\sigma}$$

$$x - b = \sigma z$$

$$dx = \sigma dz$$

$$V(x) = \int_{-\infty}^{\infty} (\sigma z)^2 \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-z2}{2}} \sigma\, dz$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \sigma^2 z^2 e^{\frac{-z2}{2}} dz$$

$$= \frac{\sigma^2}{\sqrt{2\pi}} 2\int_{0}^{\infty} z^2 e^{\frac{-z2}{2}} dz \quad \text{[Since integrand is even]}$$

Put $z^2/2 = t$

$2zdz=dt$

$2dz=dt/z=\frac{dt}{\sqrt{2t}}$

$$= \frac{2\sigma^2}{\sqrt{2\pi}} \int_{0}^{\infty} 2t\, e^{-t} \frac{dt}{\sqrt{2t}}$$

$$= \frac{2\sigma^2}{\sqrt{2\pi}} \int_{0}^{\infty} \sqrt{2t}\, e^{-t}\, dt \qquad \left(\sqrt{\ } = \gamma\right)$$

$$= \frac{2\sigma^2}{\sqrt{\pi}} \int_{0}^{\infty} e^{-t} t^{\frac{3}{2}-1} dt \left[\int_{0}^{\infty} e^{-x} x^{n-1} dx = \gamma_n\right]$$

$$= \frac{2\sigma^2}{\sqrt{\pi}} \gamma_{3/2}$$

$$= \frac{2\sigma^2}{\sqrt{\pi}} \sqrt{\frac{1}{2}+1}$$

$$= \frac{2\sigma^2}{\sqrt{\pi}} \frac{1}{2}\sqrt{\frac{1}{2}} \qquad \left[\gamma_{n+1} = n\gamma_n\right]$$

$$= \frac{\sigma^2}{\sqrt{\pi}}\sqrt{\pi} \qquad \left[\gamma^{1/2} = \sqrt{\pi}\right]$$

V(x)= σ^2

Thus the standard deviation of normal distribution is σ.

(3) Mode of Normal Distribution:

Mode is a value of x for which f(x) is maximum, i.e mode is the solution of $f^{1}(x)=0$ and $f^{11}(x)<0$ By definition, we have

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

Differentiating w .r. t 'x' we get

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

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

$$f^{1}(x) = -\left(\left(\frac{x-\mu}{\sigma}\right)\right) f(x)$$

Now $f^{1}(x)=0$

$x-\mu=0$

$x=\mu$

$$f^{11}(x) = -\left[\left(\frac{x-\mu}{\sigma^2}\right) f^{1}(x) + f(x).\frac{1}{\sigma^2}\right]$$

$$f^{11}(x) = -\left[\left(\frac{x-\mu}{\sigma^2}\right)\left(-\left(\frac{x-\mu}{\sigma^2}\right) f(x)\right) + f(x)\frac{1}{\sigma^2}\right]$$

$$= \frac{-f(x)}{\sigma^2}\left[1 - \frac{(x-\mu)^2}{\sigma^2}\right]$$

At the point $x = \mu$

$$f^{11}(x) = \frac{-f(x)}{\sigma^2}$$

$\therefore f^{11}(x) < 0$

Hence

$x = \mu$ is the mode of distribution.

4) Median of the distribution

If m is the median of the normal distribution

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

i.e $$\frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{M} e^{-1/2\left(\frac{x-\mu}{\sigma}\right)^2} dx = 1/2$$

$$\frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\mu} e^{-1/2\left(\frac{x-\mu}{\sigma}\right)^2} dx = 1/2 + \frac{1}{\sigma\sqrt{2\pi}} \int_{\mu}^{M} e^{-1/2\left(\frac{x-\mu}{\sigma}\right)^2} dx \qquad \longrightarrow (1)$$

Now

Consider $$\frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\mu} e^{-1/2\left(\frac{x-\mu}{\sigma}\right)^2} dx$$

Let $z = \frac{x-\mu}{\sigma}$

$dx = \sigma\, dz$

Upper limit $x = \mu \Longrightarrow$z=0

Lower limit $x \to -\infty \Rightarrow z \to -\infty$

$$\therefore \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\mu} e^{-1/2\left(\frac{x-\mu}{\sigma}\right)^2} dx = \frac{\sigma}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} |z| e^{\frac{-z^2}{2}} dz$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{0} e^{\frac{-z^2}{2}} \sigma \, dz$$

$$= \frac{1}{\sqrt{2\pi}} \int_{0}^{\infty} e^{\frac{-z^2}{2}} dz \ [By\ symmetric]$$

$$= \frac{1}{\sqrt{2\pi}} \sqrt{\pi/2}$$

$$= \frac{1}{2} \qquad \to (2)$$

From (1) and (2), we have

$$\frac{1}{2} + \frac{1}{\sigma\sqrt{2\pi}} \int_{dt}^{M} e^{-1/2\left(\frac{x-\mu}{\sigma}\right)^2} dx = \frac{1}{2}$$

$$\frac{1}{\sigma\sqrt{2\pi}} \int_{\mu}^{M} e^{-1/2\left(\frac{x-\mu}{\sigma}\right)^2} dx = 0$$

$\therefore \mu = M$ [If $\int_a^b f(x)dx = 0$ dx then a=b where f(x)=>0]

Hence for the normal distribution mean=median=mode.

We notice that for the normal distributions mean, median and mode coincide.

Hence, the distribution is symmetrical.

Mean deviation from the mean for normal distributions:

By definition, mean deviation about the mean.

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

$$= \int_{-\infty}^{\infty} |x-\mu| \frac{1}{\sigma\sqrt{2\pi}} e^{-1/2\left(\frac{x-\mu}{\sigma}\right)^2} dx$$

Let $z = \frac{x-\mu}{\sigma}$

$$x - \mu = \sigma z$$

$$dx = \sigma dz$$

$$= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} |\sigma z| e^{\frac{-z2}{2}} \sigma\, dz$$

$$= dz = \frac{dt}{2}$$

$$= \frac{2}{\pi} \int_{0}^{\infty} z\, e^{-t} \frac{dt}{z} \text{ [since integrand is even]}$$

$$\frac{4}{5}\sigma$$

Let $\frac{z2}{2} = t$

$$\frac{2z}{2} dz = dt$$

$$\therefore dz = \frac{dt}{z}$$

$$= \frac{2\sigma}{\sqrt{2\pi}} \int_{0}^{\infty} z\, e^{-t} \frac{dt}{z}$$

$$= \sqrt{2\pi}\ \sigma\ [0+1]$$

$$= \sqrt{\frac{2}{\pi}}\ \sigma\ [0+1]$$

$= \sigma\sqrt{\frac{2}{\pi}}$ $\left[\therefore \sqrt{\frac{2}{\pi}} = \frac{4}{5}\right]$

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

Hence the mean deviation about the mean for normal distribution is $\frac{4}{5}\sigma$

6.3 PROPERTIES OF NATURAL DISTRIBUTIONS:

(1)The graph of the normal distribution y=f(x) in the XY- Plane is known as the normal curve.

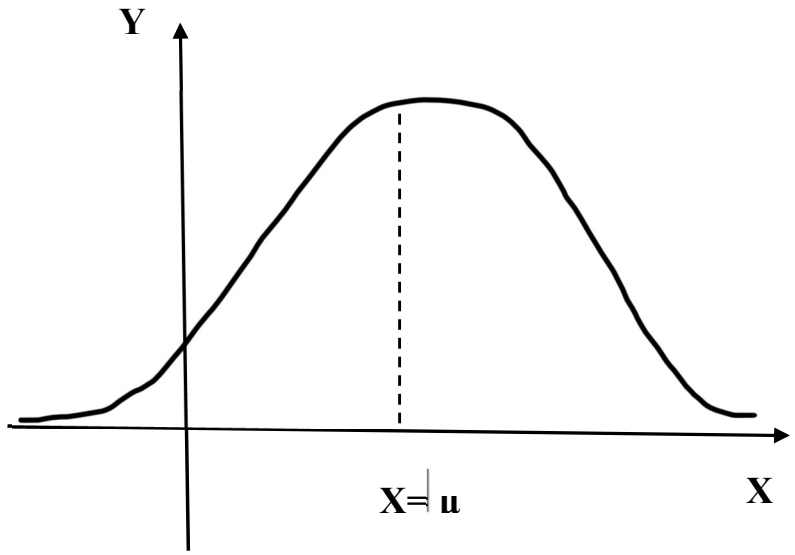


Fig 6.1 The Normal Distribution

(2) The curve is a bell-shaped curve and symmetrical with respect to mean i.e about line x = μ and the two tails on the right hand and left sides of the mean μ extends to infinity . The step of the bell is directly above the mean.

(3)Area under the normal curve represents the total population.

(4) Mean, Mode, Median of the distribution coincide at x= μ as the distribution is symmetrical. So normal curve is unimodal.

(5) X-axis is an asymmetric to the curve.

(6) Linear combination of independent normal variates is also a normal variate.

Standard normal distributions

Normal distributions with mean μ=0 and standard deviation σ=1 is known as Standard normal distributions.

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

Is called Standard normal variate.

6.4 WORKED OUT EXAMPLES

(1) If X is a normal variate with mean μ=30 and standard deviation S. Find the probabilities that (i)26 ≤ x ≤ 40 (ii)x≥45

Sol: μ=30, σ=5

Standard deviation σ=5

Standard deviation $Z = \frac{x-\mu}{\sigma}$

(i) Where x=26,

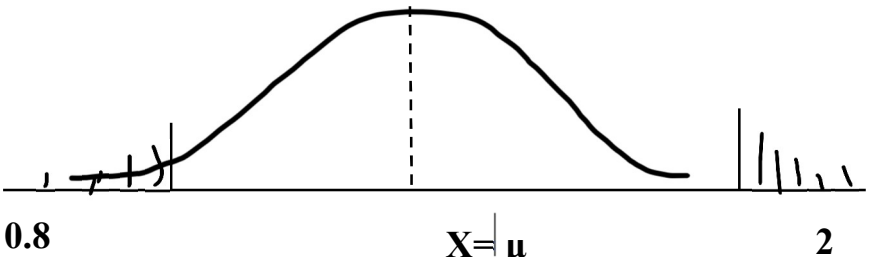
$$Z = \frac{x-\mu}{\sigma} = \frac{-4}{5} = -0.8 = Z_1$$

When x=40, $Z = \frac{40-30}{5} = \frac{10}{5} = 2 = Z_2$

$P(26 \le x \le 40) = P(Z_1 \le Z \le Z_2)$

$=P(-0.8 \le Z \le Z_2)$

$=A(2)+A(0.8)$



$P(P(26 \le x \le 40) = 0.4772+0.2881=0.7653$

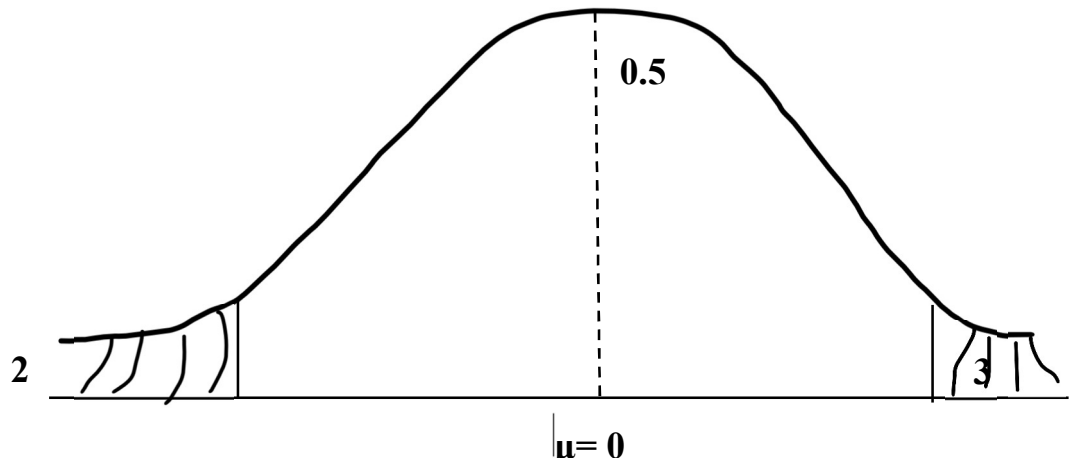
(ii) When x=45

$$Z = \frac{x-\mu}{\sigma} = \frac{45-30}{5} = \frac{15}{5} = 3$$

P(x≥45) = P(z≥3)

=0.5-A(3)

=0.5-0.4987=0.0013



2)If the masses of 300 students are normall distributed with mean(μ) 68 kgs and Standard deviation(σ) 3 kgs. How many students have masses?

(i) ≥**45 kgs**

(ii)≤64 kgs

(iii) Between 65 and 75 kgs inclusive

Sol: Let μ be the mean, σ is the standard deviation of the distribution, then

μ =68 kgs and σ =3 kgs

Let the variable x denotes the masses of students when x=72

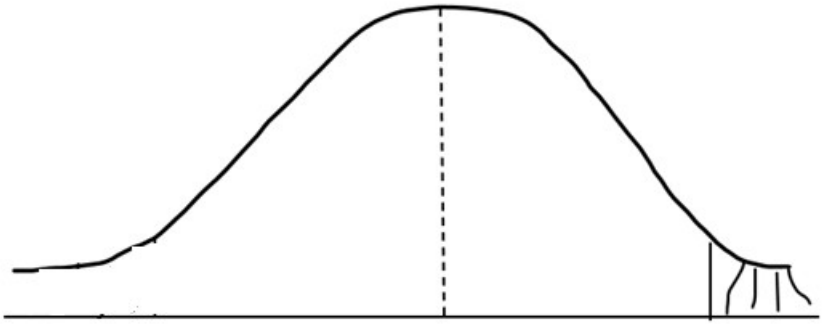
$$Z = \frac{x-\mu}{\sigma} = \frac{72-68}{3} = \frac{4}{3} = 1.33$$

P(x>72)=P(z >1.33)

=0.5-A (1.33)

=0.5-0.4082

=0.0918



(i)Number of students more than 72 kgs **1.33**

$=300\times 0.0918=28$ (Approximately)

(ii) Where x=64

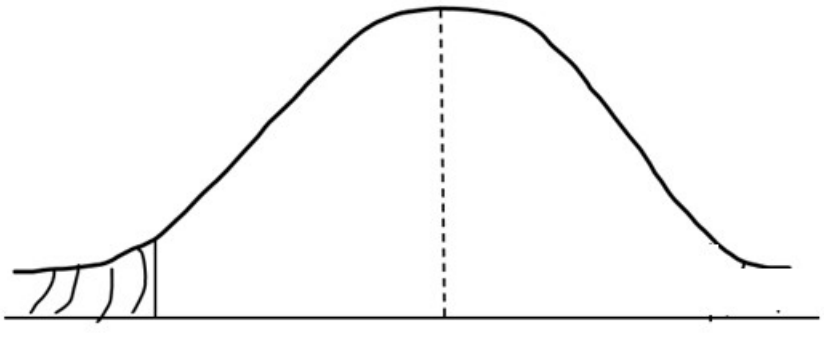
$$Z=\frac{x-\mu}{\sigma}=\frac{64-68}{3}=\frac{-4}{3}=-1.33$$

P|(x≤64)=P(2 ≤ -1.33)

=0.55-A(1.33)

=0.55-A(1.33)

=0.55-0.4082=0.0918



-1.33

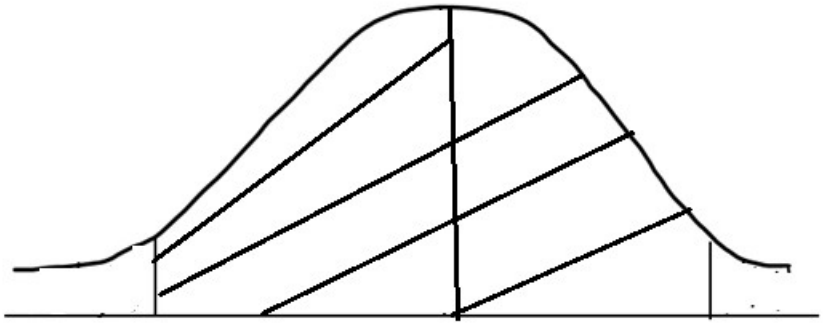
Number of students masses less than or equal to 64 kg

$=300\times 0.0918=28$

(iii) When x=65

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

$$= \frac{65-68}{3} = \frac{-3}{3} = -1$$



-1 **K=μ** **1**

When x=71

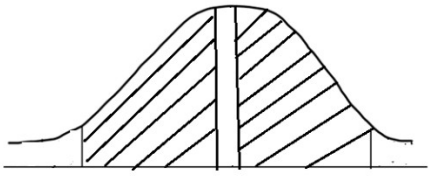
$$Z = \frac{x-\mu}{\sigma} = \frac{71-68}{3} = \frac{3}{3} = 1$$

P(65<x<71)=P(-1<Z<1)

= A(1)+A(1)

=2A(1)

=2(0.3413) =0.6826



$\Longrightarrow$Number of students whose weight lie between$(65 \leq x \leq 71)$ is $300 \times 0.6826 = 205$

3)In a Normal distribution 7% of the items under 35 and 89% are under 63. Determine the mean and variance of the distribution.

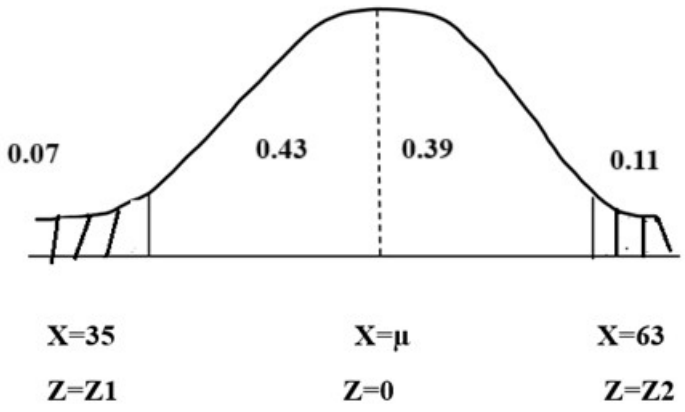
Sol: Let μ be the mean(a+z = 0) and σ be the standard deviation of the normal curve

7% of the items under 35 means the area to the left of the ordinate at x=35

Given P(x<35)=0.07

And P(x<63)=0.89

P(x>63)=1-P(x<63)=1-0.89=0.11



Where X=35, $Z = \frac{x-\mu}{\sigma} = \frac{35-\mu}{\sigma} =$ -Z_1 $\longrightarrow$ (1)

Where X=63, $Z = \frac{x-\mu}{\sigma} = \frac{63-\mu}{\sigma} = Z_2$

From the diagram,We have

($P(0<Z<Z_1)=0.39 \Longrightarrow Z_2=1.23$ [From table values]

($P(0<Z<Z_2)=0.43 \Longrightarrow Z_1=1.48$

From(1) $\frac{35-\mu}{\sigma} =$ -1.48 $\Longrightarrow 35-\mu = -1.48\sigma$

$\mu - 1.48\sigma = 35$ $\rightarrow$ (3)

From (2) $\frac{63-\mu}{\sigma} = 1.23$

$63-\mu = 1.23\,\sigma$

$\mu + 1.23\,\sigma = 63$ $\rightarrow$ (4)

On solving (3) and (4)

$\mu = 50.3 , \sigma = 10.332$

HW

4)In a normal distribution 31% of the items are under 45 and 8% are over 64. Find the mean and variance of the distribution?

5) For a normally distributed variate with mean 1 and σ 3. Find the probabilities that

(i)3.43≤x≤3

(ii)-1.43≤x≤6.19

6) Given that the mean height of students in a class is 158 cms with σ of 20 cms. Find how mean student lies between 150 cms and 170 cms. If there are 100 students in a class?

SOLUTIONS

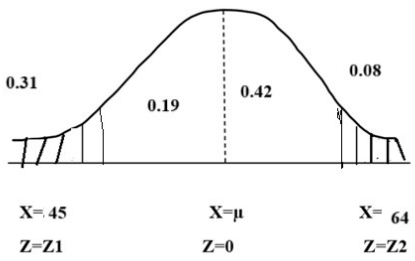
4) Sol: Let x be the continuous random variable

Let μ be the mean and σ the standard deviation

Given P(x<45)=0.31 and P(x>64)=0.08

Standard variable $Z = \frac{x - \mu}{\sigma}$

When x=45, Let Z=Z_1



So that $Z_1 = \frac{45 - \mu}{\sigma}$ $\rightarrow (1)$

$\therefore \int_{-\infty}^{\infty} \phi(Z)\,dz = 0.31$

$= \int_{-\infty}^{0} \phi(Z)\,dz - \int_{Z_1}^{0} \phi(Z)\,dz = 0.31$

$\therefore \int_{Z_1}^{0} \phi(Z)\,dz = \int_{-\infty}^{0} \phi(Z)\,dz - 0.31 = 0.5 - 0.31 = 0.19$

Hence $P(0 < Z < Z_1) = 0.19$

$\Longrightarrow Z_1 = -0.5$ (From table) $\rightarrow (2)$

When x=64, $Z = \frac{64-\mu}{\sigma} = Z_2$ (Say)

$\therefore \int_{0}^{Z_2} \phi(Z)\,dz = 0.08$ or $\int_{0}^{\infty} \phi(Z)\,dz - \int_{0}^{Z_2} \phi(Z)\,dz = 0.08$

Hence $\int_{0}^{Z_2} \phi(Z)\,dz = \int_{0}^{\infty} \phi(Z)\,dz - 0.08 = 0.5 - 0.08 = 0.42$

Thus $P(0 < Z < Z_2) = 0.42$

$Z_2 = 1.4$ (from tables) $\rightarrow (4)$

From (1) and (2),we have

$\frac{45-\mu}{\sigma} = 0.45 \quad \Rightarrow 45 - \mu = -0.5\sigma \quad \rightarrow (5)$

From (3) and (4) we get

$\frac{64-\mu}{\sigma} = 1.4 \quad \Rightarrow 64 - \mu = 1.4\sigma \quad \rightarrow (6)$

(5) - (6) gives

$(45-\mu) - 64 - \mu = -0.5\sigma - 1.4\sigma$

$\Rightarrow -19 = -1.9\sigma$

$\therefore \sigma = \frac{19}{1.9} = 10$

From (5), $\mu = 45 + 0.5\sigma = 45 + 0.5(10) = 50$

Hence mean=50 and $\sigma = 10$

5)Sol: Given $\mu = 1$ and $\sigma = 3$

(i)when x=3.43, $Z = \frac{x-\mu}{\sigma} = \frac{3.43-1}{3} = \frac{2.43}{3}$

$= 0.81 = Z_1(Say)$

When x=6.19, $Z = \frac{x-\mu}{\sigma} = \frac{6.19-1}{3} = \frac{5.19}{3}$

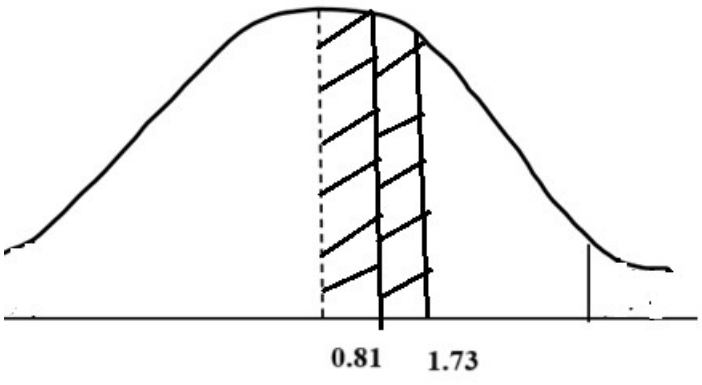
$= 1.73 = Z_2(Say)$

∴P(3.43≤x≤6.19)=P(0.81≤Z≤1.73)

=A(Z_2)-A(Z_1)

=A(1.73)-A(0.81)=0.458-0.291 (From tables)

=0.1672



(ii) When x=-1.43, $Z = \frac{x-\mu}{\sigma} = \frac{-1.43-1}{3} = -0.81 = Z_1(\text{Say})$

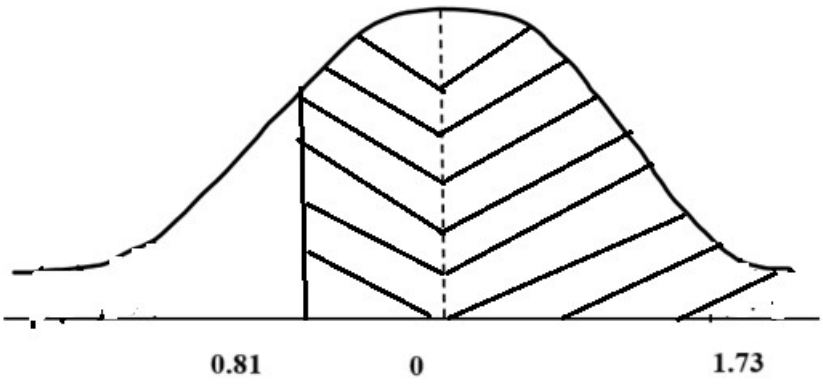
When x=6.19, $Z = \frac{x-\mu}{\sigma} = \frac{6.19-1}{3} = \frac{5.19}{3} = = 1.73 = Z_2(Say)$

$\therefore P(-1.43 \leq x \leq 6.19) = P(-0.81 \leq z \leq 1.73)$

$= A(1.73) + A(-0.81)$

$= A(1.73) + A(0.81) \quad (\therefore A(-Z) = A(Z))$

$= 0.4582 + 0.2910 = 0.7942$



6)Sol: We have

Mean, $\mu = 158$ cms and

Standard deviation $\sigma = 20$ cms

$$\therefore Z = \frac{x - \mu}{\sigma}$$

$$= \frac{x - 158}{20}$$

When x=150,

$$Z = \frac{150 - 158}{20} = \frac{-2}{5} \Rightarrow -0.4$$

When x=170,

$$Z = \frac{170 - 158}{20} = \frac{3}{5} \Rightarrow 0.6$$

$\therefore P(150 \leq x \leq 170) = P(-0.4 \leq z \leq 0.6)$

$= P(-0.4 \leq z \leq 0) + P(0 \leq z \leq 0.6)$

$= P(0 \leq z \leq 0.4) + P(0 \leq z \leq 0.6)$, [Due to symmetry]

$=0.1554+0.2257$

$=0.3811$

Number of students whose height lie between 150 cms and 170 cms

$=\text{Probability} \times \text{Total number of students} = 0.3811 \times 100 = 38$

($\therefore$Number of students should be integer)

6.5 SUMMARY

This chapter provides an in-depth introduction to the normal distribution, beginning with its formal definition and explaining how it serves as a cornerstone in statistical analysis. It outlines the key constants—such as the mean and variance—that characterize the distribution, and details its essential properties including the bell-shaped curve, symmetry about the mean, and the empirical 68-95-99.7 rule. To reinforce these concepts, the chapter includes worked-out examples that demonstrate practical applications of the normal distribution in analyzing real-world data.

6.6 TECHNICAL TERMS

- Normal Distribution
- Gaussian Distribution
- Mean
- Variance
- Standard Deviation
- Bell Curve
- Z-Score
- Probability Density Function (PDF)

6.7 SELF-ASSESSMENT QUESTIONS

SHORT:

1. What is the definition of a normal distribution?
2. Name two key constants of the normal distribution and briefly explain their significance.
3. What does the bell curve indicate about the distribution of data in a normal distribution?
4. What is the 68-95-99.7 rule in the context of a normal distribution?
5. Define a Z-score and explain its importance in the normal distribution.

ESSAY:

1. Explain the fundamental definition and properties of the normal distribution, including its symmetry, bell shape, and the empirical 68-95-99.7 rule.
2. Discuss the role of the constants—mean and variance—in shaping the normal distribution. Include how variations in these constants affect the curve's location and spread.
3. Derive the probability density function (PDF) for the normal distribution and explain the significance of each component in the formula.
4. Illustrate with worked-out examples how the normal distribution is applied in real-world data analysis, including the calculation of probabilities and Z-scores.
5. Compare and contrast the normal distribution with other common probability distributions

6.8 FURTHER READINGS

1. "An Introduction to Probability Theory and Its Applications, Volume 1" – William Feller, Wiley, 1968.
2. "Probability Theory: A Concise Course" – Y. A. Rozanov, Dover Publications, 1977.
3. "Probability Theory: An Introductory Course" – Iakov G. Sinai, Springer, 1992.
4. "Probability Theory: A Comprehensive Course" – Achim Klenke, Springer, 2020.
5. "Fat Chance: Probability from 0 to 1" – Benedict Gross, Joe Harris, Emily Riehl, Cambridge University Press, 2019.

Dr. M. Syam Sundar

Lesson - 7

WEIBULL DISTRIBUTION

OBJECTIVES:

After completion of this lesson the student should understand

a) Weibull distribution applications in reliability engineering
b) Properties of distribution
c) Inverse Weibull distributio**n**

STRUCTURE:

7.1 Introduction

7.2 Preliminaries

7.3 Weibull distribution formula

7.4 Two-Parameter Weibull Distribution

7.5 Weibull Distribution Reliability

7.6 Properties of Weibull Distribution

7.7 Inverse Weibull Distribution

7.8 Solved Problems

7.9 Facts About Weibull Distribution

7.10 Characterisation of Weibull Distribution:

7.11 Reliability Theory

7.12 Summary

7.13 Technical Terms

7.14 Self-Assessment Questions

7.15 Further Readings

7.1 INTRODUCTION

The Weibull distribution is a continuous probability distribution. It is one of the most used lifetime distributions that has applications in reliability engineering. It is an adaptable distribution that can take on the features of other kinds of distributions, depending on the value of the shape parameter. It is used to analyse the life data and helps to access the reliability of the products. In this article, we would discuss what is the Weibull distribution, what is the Weibull distribution formula, the properties, reliability, Weibull distribution

examples, two-parameter Weibull distribution, and inverse Weibull distribution in depth for your better understanding.

7.2 PRELIMINARIES

7.2.1 Definition: Weibull distribution is a type of continuous probability distribution that is used in analysing life data, times of model failure, and for accessing product reliability.

A random variable X has a Weibul distribution with three parameters c (> 0), $\alpha(> 0)$ and μ if the random variable Z =

It can also fit in a wide range of data from several other fields like hydrology, economics, biology, and many engineering sciences. It makes for an extreme value of probability distribution that is often used to model reliability, wind speeds, survival, and several other data.

The main reason for using Weibull distribution is due to its flexibility since it can simulate several other distributions just like exponential and normal distributions. Weibull distribution reliability can be measured with the help of two parameters.

Two different Weibull probability density function, also called as Weibull distribution pdf are commonly used: two-parameter pdf and three-parameter pdf.

7.3 WEIBULL DISTRIBUTION FORMULA

Let us now look at the Weibull formula.

The general expression of the Weibull pdf is noted by the three-parameter Weibull distribution expression which is given by:

β is called the shape parameter, also called as the Weibull slope

η is called the scale parameter

γ is called the location parameter

Usually, the location parameter is not much used, and you can set the value of this parameter to zero. When this is done, the pdf equation reduces to the two-parameter Weibull distribution.

7.4 TWO-PARAMETER WEIBULL DISTRIBUTION

The formula of the two-parameter Weibull distribution is practically much similar to the three-parameter Weibull distribution, the only difference being that μ isn't included:

The two-parameter Weibull is commonly used in failure analysis since no failure happens before time zero. If you know μ, the time when this failure happens, you can easily subtract it from x (i.e. time t). Hence, when you shift from the two-parameter to the three-parameter distribution, all you need to do is simply replace every instance of x with $(x - \mu)$.

7.5 WEIBULL DISTRIBUTION RELIABILITY

The Weibull distribution is commonly used in the analysis of reliability and life data since it could adapt to different situations. Depending upon the parameter values, this distribution is used for modelling a variety of behaviours for a specific function. The probability density function generally describes the distribution function. The parameters of the distribution control the location, scale, shape, of the probability density function. Many methods are used for measuring the reliability of the data. However, the Weibull distribution method is amongst the best methods for analysing the life data.

7.6 PROPERTIES OF WEIBULL DISTRIBUTION

The properties of Weibull distribution are as follows:

1. Cumulative distribution function
2. Probability density function
3. Shannon entropy
4. Moments
5. Moment generating function

7.7 INVERSE WEIBULL DISTRIBUTION

The inverse Weibull distribution could model failure rates that are much common and have applications in reliability and biological studies. A three-parameter generalized inverse Weibull distribution that has a decreasing and unimodal failure rate is presented and studied. Like the Weibull distribution, the three-parameter inverse Weibull distribution is presented for studying the different density shapes and functions of the failure rate.

The probability density function of the inverse Weibull distribution is as follows:

$$f(x)=\gamma\alpha\gamma x-(\gamma+1)\ \exp-(\alpha x)\gamma-(\alpha x)\gamma$$

7.8 SOLVED PROBLEMS

The Weibull distribution is commonly used in the analysis of reliability and life data since it is much versatile. Depending on the parameter values, the Weibull distribution is used to model several life behaviours.

7.8.1. Problem: Calculate the Weibull distribution whose α & β is 2 & 5, X1 = 1, X2 = 2.

Solution:

The first step is to substitute all these values in the above formulas.

$$P(X_1 < X < X_2) = e^{-(X_1/\beta)\alpha} - e^{-(X_2/\beta)\alpha}$$

$$P(1 < X < 2) = e^{-(1/5)2} - e^{-(2/5)2}$$

$$= 0.9608 - 0.8521$$

= 0.1087

Then calculate the mean:

Use the formula $\mu = \beta\Gamma(1 + 1/\alpha)$

= 5x $\Gamma(1+1/2)$ = 5x $\Gamma(1.5)$

= 5x 0.8864 = = 4.432

The next step is to calculate the median:

Use the formula $\beta(LN(2))^{1/\alpha}$

= $5x(0.6932)^{(1/2)}$

= 5 × 0.8326 = 4.1629

Next, calculate the variance:

Use the formula $\sigma^2 = \beta^2 [\Gamma(1 + 2/\alpha) - \Gamma(1 + 1/\alpha)^2]$

$\sigma^2 = 5^2 [\Gamma(1 + 2/2) - \Gamma(1 + 1/2)^2]$

= $25x[\Gamma(2)- \Gamma(1.5)^2]$

= 25x1−0.78571−0.7857

= 25x 0.2143

= 5.3575

Lastly, calculate the standard deviation:

σ = $\sqrt{Value\ of}$ variance

= $\sqrt{5.3575}$

= 2.3146

7.8.2 Example: Suppose the time to failure, in hours, of a bearing in a mechanical shaft, is a Weibull random variable with the following parameters. Determine the mean time until failure.

$$\alpha = \frac{1}{3} \qquad \beta = 4000$$
$$\mu = E(X) = \beta.\Gamma\left(1+\frac{1}{\alpha}\right)$$
$$\mu = 4000\Gamma\left(1+\frac{1}{1/3}\right)$$
$$\mu = 4000\Gamma(1+3)$$
$$\mu = 4000\Gamma(4)$$
$$\mu = 4000\Gamma(4-1)! = 4000.3!$$
$$\mu = 4000.6 = 24000$$

7.9 FACTS ABOUT WEIBULL DISTRIBUTION

1. The Weibull distribution can assume the characteristics of several different types of distributions. For this reason, it is extremely popular amongst the engineers and quality practitioners, who made it the commonly used distribution to model reliability data.

2. Its flexibility is the reason why engineers use the Weibull distribution for evaluating the reliability and material strengths of almost every type of things ranging from capacitors and vacuum tubes to relays and ball bearings.

3. The Weibull distribution also can model hazard functions that are increasing, decreasing, or constant, and allows it to describe any kind of phase of any item's life.

7.9.1 Weibull Distribution Parameters (Shape and Scale)

Notice that if the shape parameter(alpha) is equal to 1, then the Weibull distribution becomes the Exponential distribution! And as the scale parameter(beta) increases, the Weibull distribution becomes more symmetric.

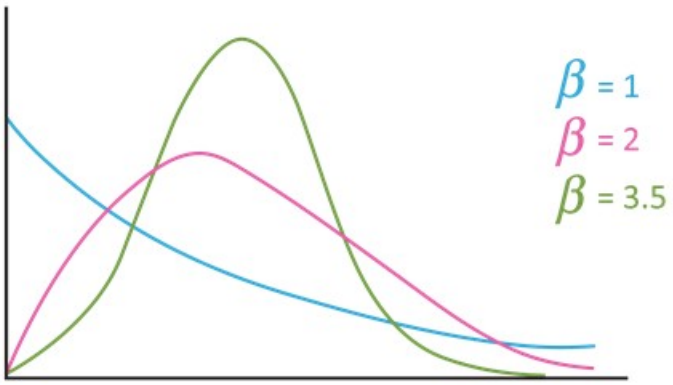


Fig 7.1 Weibull Density Curve

Mean of Weibull Distribution — Example

Then we should expect 24,000 hours until failure.

Now, using the same example, let's determine the probability that a bearing lasts a least 5000 hours.

$$P(\mathrm{X} > c) = \int f(x)dx = e^{\left(\frac{-c}{\beta}\right)^{\alpha}}$$

$$P(\mathrm{X} > 5000) = e^{\left(\frac{-5000}{4000}\right)^{\frac{1}{2}}} = 0.3405$$

7.10 CHARACTERISATION OF WEIBULL DISTRIBUTION:

Dubey, S.D. (1968) has obtained the following Result.

"Let X_i (I=1,2,….n) be i.i.d random variables. Then min (X_1,X_2….Xn) has a Weibul distribution if and only if the common distribution of X_i's is a Weibul distribution".

Proof:Let X_i (I=1,2,….n) be i.i.d *r.v* each with Weibul distribution (iii) and let Y_{min}=(X_1,X_2….Xn) Then

P(Y>y)=P[min (X_1,X_2….Xn) >y]

$$= P\left[\bigcap_{i=1}^{n} X_i > y\right]$$

$$= \prod_{i=1}^{n} P(X_i > y) = \left[P(X_i > y)\right]^n \qquad \ldots..(*)$$

$$Since \quad X_i's \quad are \quad i.i.d. \quad r.v.'s.$$

$$Now \quad P(X_i > y) = \int_y^{\infty} c\alpha^{-1}\left(\frac{x-\mu}{\alpha}\right)^{c-1} .\exp\left[-\left(\frac{x-\mu}{\alpha}\right)^c\right] dx$$

$$\int_{\frac{y-\mu}{\alpha}}^{\infty} e^{-t} dt \qquad \left[t = \left(\frac{x-\mu}{\alpha}\right)^c\right]$$

$$= \exp\left[-\left(\frac{y-\mu}{\alpha}\right)^c\right]$$

$$Substituting \quad in \quad (*), We \; get$$

$$P(Y > y) = \left[\exp\left\{\left[-\left(\frac{y-\mu}{\alpha}\right)^c\right]\right\}\right]^n$$

$$= \exp\left[-n\left(\frac{y-\mu}{\alpha}\right)^c\right]$$

$$= \exp\left[-\left\{\frac{n^{1/c}y-\mu}{\alpha}\right\}^c\right]$$

This implies that Y has the same Weibull distribution as X_i's With the difference that the parameter α is replaced by $\alpha n^{-1/c}$

7.10.1 Example: The lifetime X (in hundreds of hours) of a certain type of vacuum tube has a Weibull distribution with parameters α=2 and β=3. Compute the following:

a. E(X) and V(X)

b. P(X≤5)

c. P(1.8≤X≤5)

d. P(X≥3).

Solution:

Let X denote the lifetime(in hundreds of hours) of vaccum tube. Given that X~W (α, β) Where $\alpha = 2 \; and \; \beta = 3$

Using above formula of two parameter Weibull distribution example can be solved as below.

The probability density function of X is

$$P(X \geq 600) = 1 - P(X < 600)$$
$$= 1 - F(600)$$
$$= 1 - \left[1 - e^{-(500/300)^{0.5}}\right]$$
$$= e^{-(2)^{0.5}}$$
$$= 0.2431$$

The distribution function of X is

$$\mathrm{F}(x) = 1 - e^{-\left(\frac{x}{\beta}\right)^{\alpha}}$$

a) Mean and Variance of X

$$E(X) = \beta\,\Gamma\left(\frac{1}{\alpha}\right) + 1$$
$$= 3\,\Gamma\left(\frac{1}{2}\right) + 1$$
$$= 3\,\Gamma\left(\frac{3}{2}\right)$$
$$= 3 \times \frac{1}{2}\,\Gamma\left(\frac{1}{2}\right)$$
$$= \frac{3}{2}\sqrt{\pi}$$
$$= \frac{3}{2} \times 1.7725$$
$$= 2.6587$$

$$\mathrm{V}(X) = \beta^2\left[\Gamma\left(\frac{2}{\alpha}+1\right) - \left(\Gamma\left(\frac{1}{\alpha}\right)+1\right)^2\right]$$
$$= 3^2\left[\Gamma\left(\frac{2}{2}+1\right) - \left(\Gamma\left(\frac{1}{2}\right)+1\right)^2\right]$$
$$= 9\left[\Gamma(2) - \Gamma\left(\frac{3}{2}\right)^2\right]$$
$$= 9\left[1 - \left(\frac{1}{2}\Gamma\left(\frac{1}{2}\right)^2\right)\right]$$
$$= 9\left[1 - \left(\frac{\sqrt{\pi}}{2}\right)^2\right]$$
$$= 9\left[1 - \left(\frac{\sqrt{3.1416}}{2}\right)^2\right]$$
$$= 1.931846$$

$b) P(X \le 6)$

$$P(X \le 6) = F(6)$$
$$= 1 - e^{-(6/3)^2}$$
$$= 1 - e^{(-2)^2}$$
$$= 1 - e^{-(4)}$$
$$= 1 - 0.0183$$
$$= 0.9817$$

c) $P(1.8 \le X \le 5)$

$$P(1.8 \le X \le 5) = F(6) - \mathrm{F}(1.8)$$
$$= \left[1 - e^{-(6/3)^2}\right] - \left[1 - e^{-(1.8/3)^2}\right]$$
$$= -e^{(-0.6)^2} - e^{(-2)^2}$$
$$= e^{-(0.36)} - e^{-(4)}$$
$$= 0.6977 - 0.0183$$
$$= 0.6794$$

$d) P(X \ge 3)$

$$P(X \ge 3) = 1 - P(X < 3)$$
$$= 1 - F(3)$$
$$= 1 - \left[1 - e^{-(3/3)^2}\right]$$
$$= e^{-(1)^2}$$
$$= 0.3679$$

7.10.2 Example: Assume that the life of a packaged magnetic disk exposed to corrosive gases has a Weibull distribution with α=300 hours and β=0.5.

Calculate the probability that

a. a disk lasts at least 600 hours,

b. a disk fails before 500 hours.

Solution:

Let X denote the life of a packaged magnetic disk exposed to corrosive gases in hours.

Given that X∼W(α=300,β=0.5).

Using above formula of Two parameter Weibull distribution example can be solved as below:

The probability density function of X is

$$F(x;\alpha,\beta)=\frac{\alpha}{\beta}\left(\frac{x}{\beta}\right)^{\alpha-1} e^{-\left(\frac{x}{\beta}\right)^{\alpha}} ;\ x>0,\alpha,\beta>0$$

The distribution function of X is

$$F(x)=1-e^{-\left(\frac{x}{\beta}\right)^{\alpha}}$$

a) The probability that a disk fails before 500 hours is

$$\begin{aligned}
&P(X\le 500)=F(500)\\
&=1-e^{-(500/300)^{0.5}}\\
&=1-e^{-(1.6667)^{0.5}}\\
&=1-e^{-(1.291)}\\
&=1-0.275\\
&=0.725
\end{aligned}$$

b)The probability that a disk lasts at least 600 hours, $P(X\ge 600)$

$$\begin{aligned}
&P(X\ge 600)=1-P(X<600)\\
&=1-F(600)\\
&=1-\left[1-e^{-(500/300)^{0.5}}\right]\\
&=e^{-(2)^{0.5}}\\
&=0.2431
\end{aligned}$$

7.11 RELIABILITY THEORY

Introduction

Reliability may be defined in several ways: The idea that an item is fit for a purpose with respect to time. In the most discrete and practical sense: "Items that do not fail in use are reliable" and "Items that do fail in use are not reliable.

Reliability of a device is denoted as R(t) and defined as

$$R(t) = P[\ T > t],\ t>0.$$

Where “T” is a continuous random variable representing the life length of the device.

Some property of the reliability function *R(t):*

$$(i) R(t) = \int_{t}^{\infty} f(u)\, du$$

Where f is probability density function (p.d.f) of "T"

We assume

f(t)=0 for t<0

(ii) R(t)=-f(t)

(iii) R(t) is a non increasing function of t

$$(iv)\ R(0) = 1,\ R(\infty) = 0$$

(v)R(t)=1-F(t) Where F(t) is the cdf of T

(vi)R(t)=P(T>t)

or $R(t) = 1 - P(T \le t)$

or $R(t) = 1 - F(t)$

Hence the results

$$(iv)\ R(0) = 1,\ R(\infty) = 0$$

$$(v) \Rightarrow (iv)$$

(iii) R(t) is a non increasing function of t.

$$(iv) \Rightarrow (iii)$$

(ii) $R^1(t)$= -f(t)

$$(v) \Rightarrow (ii)$$

$$(v) \Rightarrow R(t) = 1 - F(t)$$

$$\Rightarrow R(t) = -F(t) = f(t)$$

Hence the results.

7.12 SUMMARY:

The Weibull distribution is a continuous probability distribution named after Swedish mathematician Waloddi Weibull. He originally proposed the distribution as a model for material breaking strength, but recognized the potential of the distribution in his 1951 paper

A Statistical Distribution Function of Wide Applicability. Now it's commonly used to assess product reliability, analyze life data and model failure times. The Weibull can also fit a wide range of data from many other fields, including: biology, economics, engineering sciences, and hydrology.

7.13 TECHNICAL TERMS

- Weibull Distribution
- Shape Parameter (β)
- Scale Parameter (η)
- Probability Density Function (PDF)
- Cumulative Distribution Function (CDF)
- Reliability Function (Survival Function)
- Hazard Function (Failure Rate)
- Inverse Weibull Distribution

7.14 SELF-ASSESSMENT QUESTIONS

SHORT:

1. Define the two-parameter Weibull distribution and state its probability density function (PDF).
2. What does the shape parameter (β) signify in the Weibull distribution?
3. Explain the reliability (survival) function in the context of the Weibull distribution.
4. List two key properties of the Weibull distribution that make it useful in reliability analysis.
5. What is the Inverse Weibull Distribution and how does it differ from the standard Weibull distribution?

ESSAY:

1. X_i, i = 1,2, ... ,n are independent and identically distributed (i.i.d.), Random Variable (r.v.'s) having Weibull distribution with three parameters. Show that the variable Y = min (X_1, X_2, ... , X_n) " also has Weibull distribution and identify its parameters.
2. Derive Moments of Standard Weibull Distribution?
3. The lifetime T of a device (in hours) has the Weibull distribution with shape parameter k = 1.2 and scale parameter b = 1000.
    a. Find the probability that the device will last at least 1500 hours.
    b. Approximate the mean and standard deviation of T .
    c. Compute the failure rate function.

7.15 FURTHER READING

1. Introduction to Probability and statistics by J. Susan Milton and J.C. Arnold, 4ed , TMH(2007)
2. Mathematical Statistics by R.K. Goel , Krishna Prakasan Media (P) ltd., Meerut
3. Fundamentals of Mathematical Statistics by S.C. Gupta and V.K. Kapoor, Sultan Chand & sons, New Delhi.

Dr. M. Syam Sundar

LESSON- 8

ESTIMATION

OBJECTIVE:

This Lesson is prepared in such a way that after studying the material the student is expected to have a through comprehension of the concept "Point, Interval and Central limit theorem" are the pivotal of statistical inference and analysis. The student would be equipped with theoretical as well as practical aspects of concepts.

STRUCTURE OF THE LESSON:

8.1 Introduction

8.2 Point Estimation

8.3 Interval Estimation

8.4 Properties of good estimator

8.5 Maximum Likelihood Estimation

8.6 Central limit theorem

8.7 Worked out problems

8.8 Exercise

8.9 Summary

8.10 Technical terms

8.11 Further Readings

8.1 INTRODUCTION:

In many real-life problems, the population parameter(s) is (are) unknown and someone is interested in obtaining the value(s) of parameter(s). But if the whole population is too large to study or the units of the population are destructive in nature or there is a limited resources and manpower available then it is not practically convenient to examine each unit of the population to find the value(s) of parameter(s). In such situations, one can draw sample from the population under study and utilize sample observations to estimate the parameter(s). Every one of us makes estimate(s) in our day-to-day life.

For example, a house wife estimates the monthly expenditure on the basis of particular needs, a sweet shopkeeper estimates the sale of sweets on a day, etc. So the technique of finding an estimator to produce an estimate of the unknown parameter on the basis of a sample is called estimation. There are two methods of estimation: 1. Point Estimation 2. Interval Estimation. The set of all possible values that the parameter θ or parameters $\theta_1, \theta_2, \ldots, \theta_k$ can assume is called the parameter space. It is denoted by Θ and is read as “big theta”.

For example, if parameter θ represents the average life of electric bulbs manufactured by a company, then parameter space of θ is $\Theta = \{\theta : \theta \geq 0\}$ that is, the parameter average life θ can take all possible values greater than or equal to 0.

8.2 POINT ESTIMATION

If from the observations in sample, a single value is calculated as an estimate from the unknown population parameter, the procedure is referred as point estimation, since a single point in space of all potential values is used as the estimate.

For example, in point estimation we try to estimate the μ value as a single number like $\mu = 65$ inches.

Note:

(i) A point estimate of a parameter θ is a value (based on a sample), that is a sensible guess for θ.

(ii) A point estimate is obtained by a formula ("estimator") which takes the sample data and produces an point estimate. Such formulas are called point estimators of θ.

(iii) Different samples produce different estimates though you use the same estimator, even though you use the same estimator.

For example, Consider, a luxury condominium with 702 lots. We would like to estimate the average size of the lots, their variance, as well as the proportion of lots for sale. In order to do that, a random sample with 60 lots is collected, revealing an average size of 1750 m^2 per lot, a variance of 420 m^2, and a proportion of 8% of the lots for sale.

Thus:

(a) $\bar{x}$ =1750 is a point estimate of the real population mean (μ);
(b) $S^2 = 420$ is a point estimate of the real population variance (σ^2); and
(c) $\hat{P} = 0.08$ is a point estimate of the real population proportion (p).

Again for instance, in a random sample of 80 components of a certain type, 12 are found to be defective. (a) Give a point estimate of the proportion of all not-defective units. (b) A system is to be constructed by randomly selecting two of these components and connecting them in series. Estimate the proportion of all such systems that work properly.

Solution: (a) With p denoting the true proportion of non-defective components,

$$p = (80 - 12)/ 80 = 0.85$$

(b) P_r (system works) = p^2 , since the system works if and only if both components work.

So, an estimate of this probability is $p = (68/80)^2 = 0.723$.

8.3 INTERVAL ESTIMATION

A single valued estimate or a point estimate does not in general coincide with a true value of the parameter, It is preferred to obtain a range of values or an interval. Thus, the procedure of determining an interval (a, b) that will include a population parameter, say θ with a certain probability (1 – p) is known as interval estimation. Here p is the probability that the interval does not include the true parameter value. For example, in interval estimation we try to estimate the μ value as an interval like μ is likely to fall in the interval 60 inches to 70 inches. We prefer interval estimation rather than point estimation practically.

Let p be a population of size n. Let the mean of the population be μ. Let's be a sample of the population p of size n. Let $\bar{X}$ be its mean. Now under interval estimation we will say that μ will fall in an interval around $\bar{X}$ i.e. μ will fall in the interval ($\bar{X}$ - E , $\bar{X}$ + E). Here, E is called marginal error and maximum possible error between $\bar{X}$ and μ.

Derivation of the formula for E

While estimating μ with $\bar{X}$ there always exists error which is the absolute difference of $\bar{X}$ and μ. So error E = $|\bar{X} - \mu|$. Now our aim is to minimize the error e by choosing reasonable sample size n. Suppose n > 30, for this we will take the help of central limit theorem.

Let us suppose that the probability that the error E = $|\bar{X} - \mu|$ will not exceed a prescribed limit E is 1-α.

Which implies P ($|\bar{X} - \mu| \leq E$) $= 1 - \alpha$

$$\Rightarrow P\left(\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{\pi}}} \leq \frac{E}{\frac{\sigma}{\sqrt{\pi}}}\right) = 1-\alpha$$

$\Rightarrow P(|\overline{Z}| \leq Z_{\alpha/2}) = 1-\alpha$ where $Z_{\alpha/2} = \frac{E}{\frac{\sigma}{\sqrt{\pi}}}$ and by central limit theorem $\bar{Z} = \frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{\pi}}}$

$$\Rightarrow P(-Z_{\alpha/2} \leq \overline{Z} \leq Z_{\alpha/2}) = 1-\alpha \text{ , Since } |Z| \leq a \Rightarrow -a \leq Z \leq a$$

$$\Rightarrow |A(Z_{\alpha/2}) + A(Z_{\alpha/2})| = 1-\alpha$$

$$\Rightarrow 2A(Z_{\alpha/2}) = 1-\alpha$$

$$\Rightarrow A(Z_{\alpha/2}) = \frac{1-\alpha}{2}$$

$$\Rightarrow E = Z_{\alpha/2}\frac{\sigma}{\sqrt{\pi}} = A^{-1}\left(\frac{1-\alpha}{2}\right)$$

Note:

From the above derivation we got the following formulae

1) $E = Z_{\alpha/2}\frac{\sigma}{\sqrt{\pi}}$

2) $Z_{\alpha/2} = A^{-1}\left(\frac{1-\alpha}{2}\right)$ Where n = sample size

E = maximum possible error or marginal error

$1-\alpha$ =Level of confidence or confidence with which we say that difference between X and μ will be less on than E.

Definition: $(\bar{X} - E, \bar{X} + E)$ is called confidence interval of mean.

Problem 1: Construct the confidence interval for single mean if mean of sample size of 400 is 40, standard deviation is 10 with 95% level of confidence

Solution: Sample size n = 400, Sample mean = 40, S.D $\sigma = 10$.

Given l.o.s 1-α = 95% or 1-α = 0.95

Z critical value is $Z_{\alpha/2} = A^{-1}\left(\frac{1-\alpha}{2}\right) = A^{-1}\left(\frac{0.95}{2}\right) = A^{-1}(0.475) = 1.96$

Maximum error estimate is

$$E = Z_{\alpha/2}\frac{\sigma}{\sqrt{\pi}} = \frac{1.96\times 10}{\sqrt{400}} = \frac{19.6}{20} = 0.98$$

Confidence interval of mean μ is

$(\bar{X} - E, \bar{X} + E)$ = (40-0.98, 40+0.98) = (39.2, 40.98).

conclusion: 1) the population mean μ will fall in the interval $(39.2, 41)$

2) μ will be 0.98 points away from $\bar{X}$ at the most, and

3) our confidence on this assertion is 95%

Problem 2: Calculate the confidence interval for mean if mean of sample size of 144 is 150, standard deviation is 2.

Solution: given sample size n = 144, sample mean = 150, S.D $\sigma = 2$.

Here l.o.s 1-α is not given in the data. So we fix 1-α = 100%

$\Rightarrow$ 1- α = 1

Here, $E = Z_{\alpha/2}\frac{\sigma}{\sqrt{\pi}} = A^{-1}\left(\frac{1}{2}\right) = A^{-1}\left(\frac{1}{2}\right) = A^{-1}(0.5) = 3$

$$E = Z_{\alpha/2}\frac{\sigma}{\sqrt{\pi}} = = 0.5$$

confidence interval of mean μ is

$$(\bar{X} - E, \bar{X} + E) = (150\text{-}0.5,\ 150\text{+}0.5) = (149.5,\ 150.5)$$

8.4 PROPERTIES OF GOOD ESTIMATOR:

A distinction is made between an estimate and an estimator. The numerical value of the sample mean is said to be an estimate of the population mean figure. On the other hand, the statistical measure used, that is, the method of estimation is referred to as an estimator. A good estimator, as common sense dictates, is close to the parameter being estimated. Its quality is to be evaluated in terms of the following properties: An estimator is said to be a good estimator if it is (a) Unbiased (b) Consistent (c) Efficient (d) Sufficient.

a) Unbiasedness

An estimator is said to be unbiased if its expected value is identical with the population parameter being estimated. That is if θ is an unbiased estimate of θ, then we must have E (θ) = θ. Many estimators are "Asymptotically unbiased" in the sense that the biases reduce to practically insignificant value (Zero) when n becomes sufficiently large. The estimator S_2 is an example. It should be noted that bias is estimation is not necessarily undesirable. It may turn out to be an asset in some situations.

b) Consistency

If an estimator, say θ, approaches the parameter θ closer and closer as the sample size n increases, θ is said to be a consistent estimator of θ. Stating somewhat more rigorously, the estimator θ is said is be a consistent estimator of θ if, as n approaches infinity, the probability approaches 1 that θ will differ from the parameter θ by no more than an arbitrary constant. The sample mean is an unbiased estimator of μ no matter what form the population distribution assumes, while the sample median is an unbiased estimate of μ only if the population distribution is symmetrical. The sample mean is better than the sample median as an estimate of μ in terms of both unbiasedness and consistency.

c) Efficiency

The concept of efficiency refers to the sampling variability of an estimator. If two competing estimators are both unbiased, the one with the smaller variance (for a given sample size) is said to be relatively more efficient. Stated in a somewhat different language, an estimator θ is said to be more efficient than another estimator θ_2 for θ if the variance of the first is less than the variance of the second. The smaller the variance of the estimator, the more concentrated is the distribution of the estimator around the parameter being estimated and, therefore, the better this estimations.

d) Sufficiency

An estimator is said to be sufficient if it conveys much information as is possible about the parameter which is contained in the sample. The significance of sufficiency lies in the fact that if a sufficient estimator exists, it is absolutely unnecessary to considered any other estimator; a sufficient estimator ensures that all information a sample a sample can furnished with respect to the estimation of a parameter is being utilized.

Many methods have been devised for estimating parameters that may provide estimators satisfying these properties. The two important methods are the least square method and the method of maximum likelihood.

8.5 MAXIMUM LIKELIHOOD ESTIMATION

In statistics, maximum likelihood estimation (MLE) is a method of estimating the parameters of an assumed probability distribution, given some observed data. This is achieved by maximizing a likelihood function so that, under the assumed statistical model, the observed data is most probable. The point in the parameter space that maximizes the likelihood function is called the maximum likelihood estimate.[1] The logic of maximum likelihood is both intuitive and flexible, and as such the method has become a dominant means of statistical inference.

If the likelihood function is differentiable, the derivative test for finding maxima can be applied. In some cases, the first-order conditions of the likelihood function can be solved analytically; for instance, the ordinary least squares estimator for a linear regression model maximizes the likelihood when the random errors are assumed to have normal distributions with the same variance.

Likelihood function:

Let $x_1,x_2,\ldots x_n$ be random sample of size n from a population with density function $f(x;\theta)$. Then the likelihood function of the sample values $x_1,x_2,\ldots x_n$, usually denoted by $L=L(\theta)$ is their joint density function given by $L= f(x_1;\theta), f(x_2;\theta)\ldots..,f(x_n;\theta)= \prod_{i=1}^{n} f(x_i;\theta)$. L gives the relative likelihood that the random variables a particular set of values $x_1,x_2,\ldots x_n$. For a given sample $x_1,x_2,\ldots x_n$ L becomes a function of the variable θ, the parameter.

The principle of maximum likelihood estimator consists in finding an estimator for the unknown parameter $\theta= (\theta_1,\theta_2,\ldots..,\theta_k)$ say, which maximizes the likelihood function $L(\theta)$ for variations in parameter, i.e., we wish to find $\hat{\theta}=(\theta_1, \theta_2 ,\ldots\ldots \theta_k)$ so that

$L(\hat{\theta})>L(\theta)\ \forall\ \theta \in \Theta$

i.e., $L(\hat{\theta})= \text{Sup } L(\theta), \forall\ \theta \in \Theta$

Thus if there exists a function $\hat{\theta}= \hat{\theta}(x_1,x_2,\ldots\ldots x_n)$ of the sample values which maximizes L for variations in θ, then $\hat{\theta}$ is to be taken as an estimator of $\hat{\theta}$ is usually called Maximum Likelihood Estimator (MLE). Thus $\hat{\theta}$ is the solution if any of

$$\frac{\partial L}{\partial \theta} = 0 \; and \; \frac{\partial^2 L}{\partial \theta^2} < 0$$

8.6 CENTRAL LIMIT THEOREM

The central limit theorem (CLT) is a statistical theory that states that the distribution of sample means will be normal if the sample size is large enough. This is true even if the original population distribution is not normally distributed. Which means the central limit theorem says that the sampling distribution of the mean will always be normally distributed, as long as the sample size is large enough. Regardless of whether the population has a normal, Poisson, binomial, or any other distribution, the sampling distribution of the mean will be normal.

The central limit theorem relies on the concept of a sampling distribution**,** which is the probability distribution of a statistic for a large number of samples taken from a population. Imagining an experiment may help you to understand sampling distributions:

- Suppose that you draw a random sample from a population and calculate a statistic for the sample, such as the mean.
- Now you draw another random sample of the same size, and again calculate the mean.
- You repeat this process many times, and end up with a large number of means, one for each sample.
- The distribution of the sample means is an example of a sampling distribution.
- A normal distribution is a symmetrical, bell-shaped distribution, with increasingly fewer observations the further from the center of the distribution.
- Fortunately, you don't need to actually repeatedly sample a population to know the shape of the sampling distribution. The parameters of the sampling distribution of the mean are determined by the parameters of the population:
- The mean of the sampling distribution is the mean of the population.
- The mean of the sampling distribution is the mean of the population.

$$\mu_{\overline{x}} = \mu$$

- The standard deviation of the sampling distribution is the standard deviation of the population divided by the square root of the sample size.

$$\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}}$$

We can describe the sampling distribution of the mean using this notation:

$$\overline{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

Where:

- $\bar{X}$ is the sampling distribution of the sample means
- ~ means "follows the distribution"
- N is the normal distribution
- μ is the mean of the population
- σ is the standard deviation of the population
- n is the sample size

- The mean of the sampling distribution is the mean of the population. $\mu_{\bar{x}} = \mu$

Sample size and the central limit theorem

The **sample size** (n) is the number of observations drawn from the population for each sample. The sample size is the same for all samples.

The sample size affects the sampling distribution of the mean in two ways.

Sample size and normality

The larger the sample size, the more closely the sampling distribution will follow a normal distribution. When the sample size is small, the sampling distribution of the mean is sometimes non-normal. That's because the central limit theorem only holds true when the sample size is "sufficiently large. " By convention, we consider a sample size of 30 to be "sufficiently large."

- **When $n < 30$**, the central limit theorem doesn't apply. The sampling distribution will follow a similar distribution to the population. Therefore, the sampling distribution will only be normal if the population is normal.

- **When $n \geq 30$**, the central limit theorem applies. The sampling distribution will approximately follow a normal distribution.

Sample size and standard deviations

The sample size affects the standard deviation of the sampling distribution. Standard deviation is a measure of the variability or spread of the distribution (i.e., how wide or narrow it is).

- **When *n* is low**, the standard deviation is high. There's a lot of spread in the samples' means because they aren't precise estimates of the population's mean.

- **When *n* is high**, the standard deviation is low. There's not much spread in the samples' means because they're precise estimates of the population's mean.

Conditions of the central limit theorem

The central limit theorem states that the sampling distribution of the mean will always follow a normal distribution under the following conditions:

(a) The sample size is **sufficiently large**. This condition is usually met if the sample size is $n \geq 30$.
(b) The samples are **independent and identically distributed (i.i.d.) random variables**. This condition is usually met if the sampling is random.

(c) The population's distribution has **finite variance**. Central limit theorem doesn't apply to distributions with infinite variance, such as the Cauchy distribution. Most distributions have finite variance.

Importance of the central limit theorem

(a) The central limit theorem is one of the most fundamental statistical theorems. In fact, the "central" in "central limit theorem" refers to the importance of the theorem.

(b) **Parametric tests**, such as *t* tests, ANOVAs, and linear regression, have more statistical power than most non-parametric tests. Their statistical power comes from assumptions about populations' distributions that are based on the central limit theorem.

(c) Central limit theorem examples

Applying the central limit theorem to real distributions may help you to better understand how it works.

(d) Continuous distribution

(e) Suppose that you're interested in the age that people retire in the United States. The **population** is all retired Americans, and the distribution of the population might look something like this:

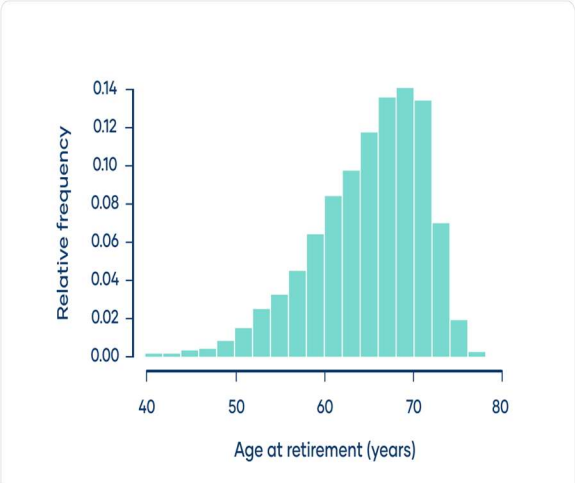


Fig 8.1 Continuous distribution: Example

Age at retirement follows a left-skewed distribution. Most people retire within about five years of the mean retirement age of 65 years. However, there's a "long tail" of people who retire much younger, such as at 50 or even 40 years old. The population has a standard deviation of 6 years. Imagine that you take a small sample of the population. You randomly select five retirees and ask them what age they are retired.

Example: Central limit theorem; sample of $n = 5$

68 73 70 62 63

The mean of the sample is an estimate of the population mean. It might not be a very precise estimate, since the sample size is only 5. Central limit theorem: mean of a small sample mean

$= (68 + 73 + 70 + 62 + 63) / 5$

$\Rightarrow$ mean = 67.2 years

Suppose that you repeat this procedure 10 times, taking samples of five retirees, and calculating the mean of each sample. This is a sampling distribution of the mean.

Example: Central limit theorem; means of 10 small samples

60.8 57.8 62.2 68.6 67.4 67.8 68.3 66.5 62.1 65.6

If you repeat the procedure many more times, a histogram of the sample means will look something like this:

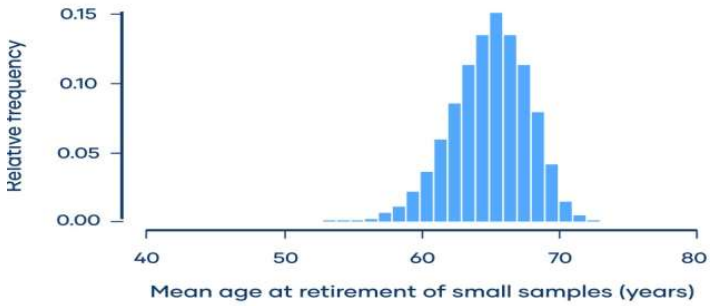


Fig 8.2 Central limit theorem : Example

Although this sampling distribution is more normally distributed than the population, it still has a bit of a left skew. Notice also that the spread of the sampling distribution is less than the spread of the population.

The **central limit theorem** says that the sampling distribution of the mean will always follow a normal distribution when the sample size is sufficiently large. This sampling distribution of the mean isn't normally distributed because its sample size isn't sufficiently large. Now,

imagine that you take a large sample of the population. You randomly select 50 retirees and ask them what age they retired.

Example: Central limit theorem; sample of $n = 50$

73	49	62	68	72	71	65	60	69	61
62	75	66	63	66	68	76	68	54	74
68	60	72	63	57	64	65	59	72	52
52	72	69	62	68	64	60	65	53	69
59	68	67	71	69	70	52	62	64	68

The mean of the sample is an estimate of the population mean. It's a precise estimate, because the sample size is large.

Example: Central limit theorem; mean of a large sample mean = 64.8 years

Again, you can repeat this procedure many more times, taking samples of fifty retirees, and calculating the mean of each sample:

In the histogram, you can see that this sampling distribution is normally distributed, as predicted by the central limit theorem.

The standard deviation of this sampling distribution is 0.85 years, which is less than the spread of the small sample sampling distribution, and much less than the spread of the population. If you were to increase the sample size further, the spread would decrease even more.

We can use the central limit theorem formula to describe the sampling distribution:

$$\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

$$\mu = 65,\ \sigma = 6,\ n = 50$$

$$\bar{X} \sim N\left(65, \frac{6}{\sqrt{50}}\right)$$

$$\bar{X} \sim N(65, 0.85)$$

In case of Discrete distribution

Approximately 10% of people are left-handed. If we assign a value of 1 to left-handedness and a value of 0 to right-handedness, the probability distribution of left-handedness for the **population** of all humans looks like this:

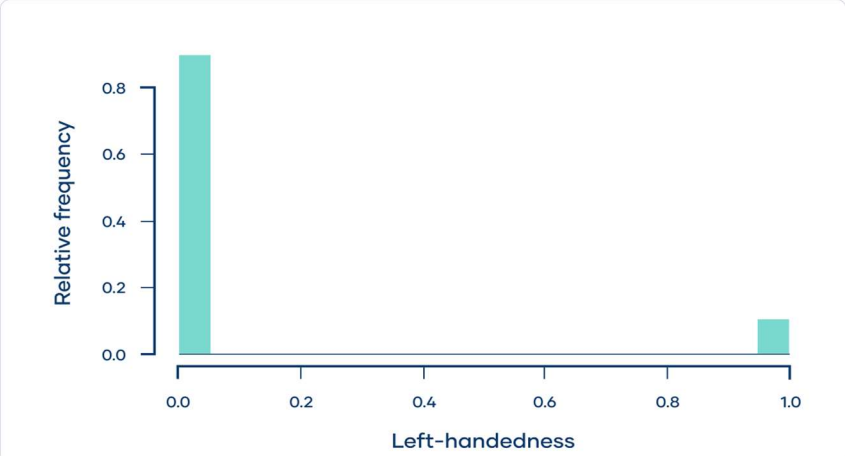


Fig 8.3 Discrete distribution: Example

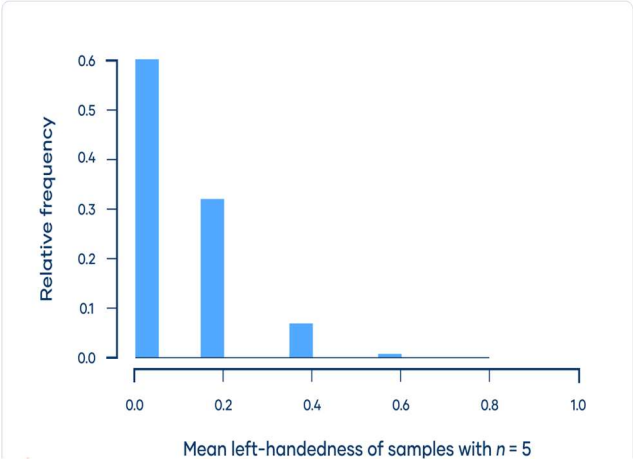
The population mean is the proportion of people who are left-handed (0.1). The population standard deviation is 0.3.

Imagine that you take a random sample of five people and ask them whether they're left-handed.

For an example of Central limit theorem; sample of n = 5, 0 0 0 1 0
The mean of the sample is an estimate of the population mean. It might not be a very precise estimate, since the sample size is only 5.
Central limit theorem; mean of a small sample mean = (0 + 0 + 0 + 1 + 0) / 5

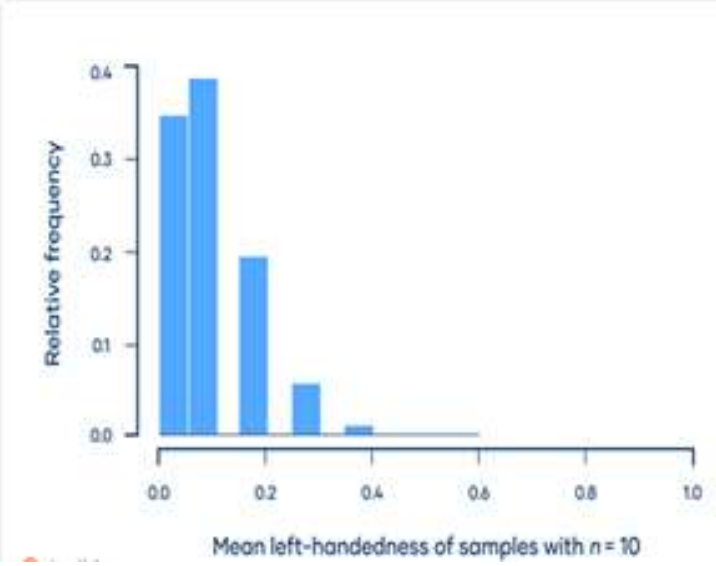
mean = 0.2
Imagine you repeat this process 10 times, randomly sampling five people and calculating the mean of the sample. This is a **sampling distribution of the mean**.

For example, in Central limit theorem; the means of 10 small samples
0,0,0.4,0.2,0.2,0,0.4,0
If you repeat this process many more times, the distribution will look something like this:



The sampling distribution isn't normally distributed because the sample size isn't sufficiently large for the central limit theorem to apply. As the sample size increases, the sampling distribution looks increasingly similar to a normal distribution, and the spread decreases:

***n* = 10**, ***n* = 20**, ***n* = 30**, ***n* = 100**



he sampling distribution of the mean for samples with $n = 30$ approaches normality. When the sample size is increased further to $n = 100$, the sampling distribution follows a normal distribution. We can use the central limit theorem formula to describe the sampling distribution for $n = 100$.

$$\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$$

$\mu = 0.1$, $\sigma = 0.3$, $n = 100$

$$\bar{X} \sim N(0.1, \frac{0.3}{\sqrt{100}})$$

$$\bar{X} \sim N(0.1, 0.03)$$

8.7 WORKED OUT EXAMPLES:

Example 1. Let X be distributed in the Poisson form with parameter θ. Show that the only unbiased estimator of $e^{-(k+1)\theta}, k > 0$ *is* $T(x) = (-k)^x$, so that T(x)>0, if x is even and T(x)<0, if x is odd.

Solution: Now $E\{T(x)\} = [E(-k)^x] = \sum_{x=0}^{\infty}(-(k)^x).e^{-\theta}.\theta^x/x!$

$$= e^{-\theta}\sum_{x=0}^{\infty}(-(k\theta)^x)/x!$$

$$= e^{-\theta}.e^{-k\theta}$$

$$= e^{-(1+k)\theta}$$

$\therefore T(x) = (-k)^x$ is an unbiased estimate of $e^{-(k+1)\theta}, k > 0$.

Example 2. Show that in a random sampling from a normal population, the sample mean is a consistent estimator for the population mean.

Solution: Let $x_1, x_2, \ldots\ldots x_n$ be the sample drawn from the normal population with mean μ and finite variance σ^2. From the result that the statistic $z = (\bar{x} - \mu)\varepsilon\frac{\sqrt{n}}{q}$ is a standard normal variate.

$$\therefore P[(\bar{x} - \mu) < \varepsilon] = [P|z| < \varepsilon\frac{\sqrt{n}}{q}] \int_{\frac{-\varepsilon\sqrt{n}}{\sigma}}^{\frac{+\varepsilon\sqrt{n}}{\sigma}} \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}} dz$$

Hence, given any number η, however small, we can choose n, so that area under the standard normal curve between $\frac{-\varepsilon\sqrt{n}}{\sigma}$ and $\frac{+\varepsilon\sqrt{n}}{\sigma}$ becomes greater than 1-η.

Example 3. Show that, in estimating for the mean for random sampling from a normal population, sample mean is more efficient than the sample median. Discuss also the populations of the two estimators for Cauchy population.

Solution: For a normal population with finite variance σ^2, the sample mean and sample variance are consistent estimators for the population mean and population variance respectively.

For any n, Var(mean) = σ^2/n and for large n, Var(median)= $\pi\,\sigma^2/2n$.

Since var(mean) is smaller then var(median), the mean is more efficient than the median for large n. For small n, by actual calculations it can be proved that var(mean) is smaller than var(median). Also the efficiency of the median relative to the mean is $2/\pi$, i.e., 0.637.

For Cauchy population, a direct comparison is not possible since the mean is not even a consistent estimator. However, var(median)= $\pi^2/4n$ and obviously in this case, median is a better estimator than mean.

Example 4: Let $x_1, x_2, \ldots\ldots x_n$ be random sample from Cauchy population

$f(x;\theta) = \dfrac{1}{\pi(1+(x-\theta)^2}$,$-\infty<x<\infty$, $-\infty<\theta<\infty$. Does there exist a sufficient statistic for θ?

Solution: $x_1, x_2, \ldots\ldots x_n$ be a random sample from Cauchy population whose pdf is

$f(x;\theta) = \dfrac{1}{\pi(1+(x-\theta)^2}$, $-\infty<x<\infty$, $-\infty<\theta<\infty$

the likelihood function is $L = \prod_{i=1}^{n} f(x_i,\theta) = \dfrac{1}{\pi^n}\prod_{i=1}^{n}\left[\dfrac{1}{1+(x_i-\theta)^2}\right]$

Now L cannot be factorised, such that one factor does not contain θ. Hence there does not exist a sufficient estimator for θ.

Example 5: A soft-drink vending machine is set so that the amount of drink dispensed is a random variable with a mean of 200 milliliters and a standard deviation of 15 milliliters. What is the probability that the average (mean) amount dispensed in a random of size 36 is at least 204 milliliters?

Solution: The distribution of $\bar{x}$ has the mean μ_x=200 and the standard deviation $\sigma_x = \dfrac{15}{\sqrt{36}}$ =2.5, and according to the central limit theorem, this distribution is approximately normal.

Since, Z= $\dfrac{\bar{x}-\mu}{\sigma} = \dfrac{204-200}{2.5} = 1.6$ which follows from normal tables that

$P(x \geq 204) = P(Z \geq 1.6) = 0.5000 - 0.4452 = 0.0548$.

Hence it is to be noted that when the population we are sampling is normal, the distribution of $\bar{x}$ is a normal distribution regardless of the size of n.

Example 6: Given x successes in n trials, find the maximum likelihood estimate of the parameter θ of the corresponding binomial distribution.

Solution: To find the value of θ which maximizes $L(\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}$

It will be convenient to make use of the fact that the value of θ which maximizes L(θ) will also maximize

$$LnL(\theta) = Ln\binom{n}{x} + x.\ln\theta + (n-x).\ln(1-\theta)$$

$$\frac{d[\ln L(\theta)]}{d\theta} = \frac{x}{\theta} = \frac{n-x}{1-\theta}$$

And equating this derivative to 0 and solving for θ, we find that the likelihood function has a maximum at θ=$\frac{x}{n}$. This is the maximum likelihood estimate of the binomial parameter θ, and we refer to $\hat{\theta} = \frac{x}{n}$ as the corresponding maximum likelihood estimator.

Example 7: Let $x_1, x_2, \ldots x_n$ are the values of a random sample from an exponential population, find the maximum likelihood estimator of its parameter θ.

Solution: Since the likelihood function is given by $L(\theta)=f(x_1,x_2,\ldots.x_n; \theta)$

the likelihood function is $L = \prod_{i=1}^{n} f(x_i, \theta)$

$$= \left(\frac{1}{\theta}\right)^n = e^{-1/\theta \sum_{i=1}^{n} x_i}$$

Differentiation of L(θ) w.r.t `θ` yields

$$\frac{d[\ln L(\theta)]}{d\theta} = \frac{n}{\theta} + \frac{1}{\theta^2} = \sum_{i=1}^{n} x_i$$

equating this derivative to 0 and solving for θ, we get maximum likelihood estimate

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

Hence, maximum likelihood estimate of θ is $\hat{\theta} = \bar{x}$

8.8 EXERCISE

1) Show that, for a random sampling from a normal population, in estimating for the mean, the estimator $x_1$the first observation of the sample is unbiased but not consistent?
2) Show that the sample variance is a consistent estimator for the population variance of a normal distribution?
3) Show that for samples of size n from normal population with mean μ and variance σ^2, the statistic $\hat{\mu}=\sum_{i=1}^{n} x_i/n+1$ is most efficient for estimation μ, though it is biased.
4) Let $x_1,x_2,\ldots x_n$ be a random sample from a uniform population on [0, θ], Find a sufficient estimator for θ.
5) Calculate the confidence interval for mean if mean of sample size of 81 is 50, standard deviation is 3 at 99% of confidence.
6) For a random sampling from a normal population, find the maximum likelihood estimators for a) the population mean, when the population variance is known.
   b) the population variance, when the population mean is known
   c) The simultaneous estimation of both the population mean and variance.
7) Show that the rectangular population with $f(x,\theta)= 1/\theta$, $0\leq x\leq\theta$, $0<\theta<\infty$, =0 elsewhere, the maximum likelihood estimator for θ is the largest member of the sample.

8) Find the mle for θ for $f(x)=\theta x^{\theta}$, $0<x<1$?

9) What is the mle for α in the density $f(x)=(\alpha+1) x^{\alpha}$, $0<x<1$?

10) Find the estimate for α in samples of size 2 when the density function is

$f(x, \alpha) = 2.(\alpha-x)/\alpha^2$, $0<x<\alpha$.

11) Let X be the height of a randomly chosen individual from a population. In order to estimate the mean and variance of X, we observe a random sample X1,X2,⋯⋯,X7. Thus, Xi's are i.i.d. and have the same distribution as X. We obtain the following values (in centimeters):166.8, 171. 4, 169.1, 178.5, 168.0, 157.9, 170.1 166.8, 171.4, 169.1, 178.5, 168.0, 157.9, 170.1. Find the values of the sample mean, the sample variance, and the sample standard deviation for the observed sample.

8.9 SUMMARY:

In this lesson an attempt is made to explain the concepts of Estimation- Point, Interval, Maximum likelihood estimation and Central limit theorem associated with them

along with theory and practical. A number of examples are worked out and a good number of exercises are also given.

8.10 TECHNICAL TERMS:

- Point Estimation
- Interval Estimation
- Properties of good estimator
- Maximum Likelihood Estimation
- Central limit Theorem

8.11 Further Reading

1. Introduction to Probability and statistics by J. Susan Milton and J.C. Arnold, 4ed , TMH(2007)
2. Mathematical Statistics by R.K. Goel , Krishna Prakasan Media (P) ltd., Meerut
3. Fundamentals of Mathematical Statistics by S.C. Gupta and V.K. Kapoor, Sultan Chand & sons, New Delhi.

Dr. M. Syam Sundar

LESSON 9

INFERENCE ON THE MEAN

OBJECTIVES:

After completion of this lesson the Student should able to understand

→Testing of Hypothesis on mean

⟶Significance of testing on mean

→Hypothesis and significance test on the mean

STRUCTURE OF THE LESSON

9.1 Introduction

9.2 Null hypothesis & Alternative hypotheses

9.3 Simple and Composite hypothesis

9.4 Type-I and Type-II Errors

9.5 Test of Significance

9.6 Level of Significance

9.7. Test for Single mean

9.8 Solved Problems

9.9 Summary

9.10 Technical Terms

9.11 Exercises

9.12 Further Reading

9.1 INTRODUCTION

In tentative Statement about distribution of one or more random Variables, the purpose is to Satisfy some parametric of the population. Then we Shall apply certain tests to direct whether a pre assigned value of the parameter is acceptable in the light of observation in the Sample.

This Process is known as “Testing of Hypothesis”.

There are two types of Hypothesis

(a) Null hypothesis

(b) Composite hypothesis.

9.2 NULL HYPOTHESES & ALTERNATIVE HYPOTHESIS

A hypothesis test is a binary question about data distribution.

9.2.1 Definition: A null hypothesis is a hypothesis which is tested for possible rejection under the assumption that it is true. It is so to say, " a straw man" that we Set up, possible for the purpose of knocking down. It is denoted by H_0.

9.2.2 Note: By accepting a null hypothesis, we do not mean that it is proved to be true. This only implies that based on the statistics Calculated from the sample, we find no reason to question the validity of the hypothesis.

9.2.3 Example

For example, the hypothesis may be put in a form "Paddy variety A will give the Same yield per hectare as that of Variety B' (or) there is no difference between the average yields of paddy varieties A and B. These hypotheses are in definite terms.

Thus these hypotheses form a basis to work. So, such working hypothesis is known as null hypothesis.

9.2.4 Definition:

A statistical hypothesis used in hypothesis testing which states that there is a Significant difference between the set of variables, It is often referred to as the hypotheses other than the null hypothesis is called an alternative hypothesis. It is denoted by H_1.

Note that the acceptance of alternative hypothesis depends on the rejection of the null hypothesis. That is, until and unless null hypothesis rejected, an alternative hypothesis cannot be accepted.

9.2.5 Example: A Coin was tossed 400 times and the head turned up 216 times. Write Null hypothesis and alternative hypothesis the coin is unbiased.

Solution: Null hypothesis

H_0: The coin is unbiased, i.e P=0.5

Alternative hypothesis

H_1: The Coin is not unbiased (biased):P=0.5

9.2.6 Key differences between Null and Alternative hypothesis:

Here we discussed some important points of difference between Null and Alternative hypothesis.

1) A null hypothesis is a statement in which there is no relationship between two variables where as an Alternative hypothesis is a statement, that is simply the inverse of the null hypothesis.

i.e there is some statistical significance between two measured Phenomenon.

2) A null hypothesis is what, anybody tries to disprove where as an alternative hypothesis is what he wants to prove

3)A null hypothesis represents, no observed effect where as an alternative hypothesis reflects some observed effect.

4)If the null hypothesis is accepted, no changes will be made in the opinions of actions. Conversely, of the alternative hypothesis is accepted, it will result in the changes in the opinions or actions.

5)In null hypothesis, the observations are the outcome of chance where as in alternative hypothesis, the observations are an outcome of real effect.

Note:

It is observed that there are two outcomes of a Statistical test, that is, first a null hypothesis is rejected and alternative hypothesis is accepted, Second, null hypothesis is accepted or the basis of the evidence, In simple terms, a null hypothesis is just opposite of alternative hypothesis.

Example: The average age of online consumers of a few years ago was 23.3 years. As older individuals gain confidence with the internet. It is belied that the average age has increased. State the hypothesis being tested.

Solution: In the given problem,

μ = average age of on line Consumers and is the unknown parameter.

By this, the hypothesis value, μ_0 is 23.3 years and we want to test whether μ is in fact more than this.

Hence We have

null hypothesis, H_0: $\mu = 23.3$

Alternative hypothesis, H_1: $\mu > 23.3$

This shows that the null hypothesis is capturing the statement of "no effect" and the average age is still 23.3 years

The Alternative captures, the statement of interest which is that the average age is more than 23.3 years.

9.3 SIMPLE AND COMPOSITE HYPOTHESIS

9.3.1. Definition: If a hypothesis specifies only one value or exact value of the population parameter then it is known as simple hypothesis.

If a hypothesis specifies not just one value but a range of values that the population parameter may assume is called a composite hypothesis.

9.3.2 Example: Consider a) A Customer of motorcycle wants to test whether the claim of motor cycle of certain brand gives the average milage 60km/liter is true (or) false

b)The businessman of banana wants to test whether the average weight of a banana of kerala is more than 200 gms.

Here in a) $\mu = 60$. km |Lt is simple hypothesis because it gives a single value of parameter $\mu = 60$

In b) $\mu > 200$ gms is composite hypothesis because it does not specify the exact average value of weight of a banana. So, it may be 250,300,…

9.3.3 Example:

If $x_1, x_2 \ldots x_n$ is a random sample of size 'n' from a normal population with mean 'μ′ and Variance σ^2. Then the hypothesis

$$H_0 : \mu = \mu_0 , \ \sigma^2 = \sigma_0^2$$

Is a sample hypothesis, where as

The following hypothesis is a composite hypothesis:

$a) \mu = \mu_0$

$b) \sigma^2 = \sigma_0^2$

$c) \mu < \mu_0 , \sigma^2 = \sigma_0^2$

$d) \mu > \mu_0 , \sigma^2 = \sigma_0^2$

$e) \mu = \mu_0 , \sigma^2 < \sigma_0^2$

$f) \mu = \mu_0 , \sigma^2 > \sigma_0^2$

$g) \mu < \mu_0 , \sigma^2 > \sigma_0^2$

9.3.4 Note: A hypothesis which does not specify completely 'r' parameters of a population is termed as a Composite hypothesis with 'r' degree of freedom.

9.3.5 The best test for a simple hypothesis:

Often the test criterion it to be determined by controlling α & β where 'α' is the probability rejecting H_0 and 'β' is the probability of H_0 of accepting H_0.

The ideal thing would be a minimize 'α' and 'β' simultaneously but in practice when 'α' is minimized, β becomes large and vice versa. Hence the attempt is to minimize β for a fixed 'α' and if there exists such a test criterion, it is called the "best test' for a simple hypothesis.

9.3.6 Critical Region:

The basis of the testing of hypothesis is the division of the sample space into two exclusive regions, The region of acceptance and region of rejection.

If the sample point falls in the region of rejection, H_0 is rejected.

The region of rejection is called "CRITICAL REGION".

The probability that the sample point falls in the critical region is called the Size of the Critical region of the test.

similarly the probability of a sample point falling in the region of acceptance is β.

The two generally accepted levels of rejections are 5% and l%. In 5% level of significance the Confidence interval is 95%.

Usually we take two critical regions which Cover 5%. and 1% areas of the normal curve

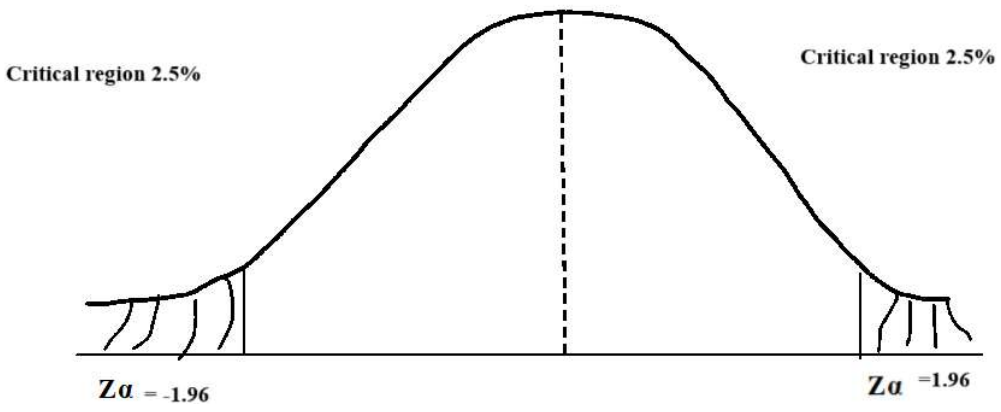


Fig 9.1 Critical Region

9.4 Type -I and Typ-II errors

9.4.1 Definition:

When a hypothesis H_0 is tested against an alternative H_1, usually there can arise one of the two types of errors. when a null hypothesis H_0 is rejected, although it is true, this is known as a TYPE-I error.

On the other hand, the null hypothesis H_0 is accepted when it is false, i known as type-II error.

Note that here we will assume that rejection of H_0 amounts to acceptance of H_1.

9.4.2 Example: for example, if

H_0: μ=20 and H_1: μ=30 in $N(\mu,\sigma^2)$ the rejection of H_0 is equivalent to the acceptance of μ=30.

9.4.3 Definition: The probabilities of committing the type I and type II errors are called Sizes of errors and denoted by α and β.

9.4.4 Example: An Engineer infers that a packet of Screws is Sub standard when actually it is not. It is an error caused due to poor or appropriate (faculty)sample.

Similarly, a packet of Screws may infer good when actually it is sub standard.

So, here we can Commit two kinds of errors while testing a hypothesis which are summarized in the following table:

Decision	H_0 true	H_1 true
Reject H_0	Type-I error	Correct Decision
Do not Reject H_0	Correct Decision	Type-II Error

9.4.5 Example: For testing H_0 :θ=1 against H_1:θ=2, The probability density function of the variable is given by

$$=f(x,\theta)=\begin{cases}\frac{1}{\theta} & 0\le x\le\theta\\ 0 & elsewhere\end{cases}$$

Obtain type-I and type-II errors when critical region is X≥0.4. Also obtain power function of the test.

Solution: Given probability density function(pdf) is

$$f(x,\theta)=\begin{cases}\frac{1}{\theta} & 0\le x\le\theta \\ 0 & elsewhere\end{cases} \qquad \rightarrow(1)$$

Here we have rejection and non rejection regions as

$$\omega=\{X: X\ge 0.4\} \text{ and}$$
$$\overline{\omega}=\{X: X<0.4\}$$

We must test the null hypothesis

$H_{0:}\ \theta=1$ against $H_{1:}\ \theta=2$

So, the size of type-1 error is given by

$$\alpha=P[X\in\omega/H_0]=P\{X\ge 0.4/\theta=1\}$$
$$=\left[\int_{0.4}^{\theta} f(x,\theta)\,dx\right]_{\theta=1} \qquad \rightarrow(2)$$
$$(Since \quad P(x\ge a)=\int_a^{\infty} f(x,\theta)\,dx)$$

Now by using

$$f(x,\theta)=\frac{1}{\theta},\ 0\le x\le\theta,$$

We get from equation (2),

$$\alpha=\left[\int_{0.4}^{\theta}\frac{1}{\theta}dx\right]_{\theta=1}=\int_{0.4}^{1}dx=(x)_{0.4}^{1}$$

$$=1-0.4=0.6$$

Similarly, the size of type -II error is given by

$$\beta=P[X\in\overline{\omega}/H_1]=P[X<0.4/\theta=2]$$
$$=P\left[\int_0^{0.4}\frac{1}{\theta}dx\right]_{\theta=2}=\int_0^{0.4}\frac{1}{2}dx$$
$$=\frac{1}{2}(x)_0^{0.4}=\frac{1}{2}(0.4-0)=\frac{0.4}{2}=0.2$$

Here the power function of the test

$1-\beta=1-0.2=0.8$

9.5 TEST OF SIGNIFICANCE

A very important aspect of the Sampling theory is the study of the tests of Significance, which enable us to decide on the basis of the sample results, if

a) the derivation between the observed sample statistic and the hypothetical parameter value (or)

(b) the derivation between two independent Sample Statistics, is significant (or) might be attributed to chance (or) the fluctuations of Sampling.

9.5.1 Tests of Significance for large samples:

Here we will discuss the tests of significance when Samples are large. We have seen that for large values of 'n', the number of trails, almost all the distributions, for examples binomial, Poisson, negative binomial etc, are very Closely approximated by normal distribution. In this case we apply the normal test which is based upon the following fundamental property of the normal probability Curve.

If

$$X \sim N(\mu, \sigma^2) \text{ then}$$

$$Z = \frac{X-\mu}{\sigma} = \frac{X - E(\mathrm{X})}{\sqrt{V(\mathrm{x})}} \sim N(0,1)$$

Procedure of deciding whether population parameter is correct or not is called Test of Significance.

9.5.2 Note:

Test of Significance involve following ratio

Standard Score = $\frac{\text{observation} - \text{mean}}{Standard\ deviation}$

Where, for population proportion 'p' in particular,

observation = p mean= p

and $S.D = \sqrt{\frac{p(1-p)}{n}}$

And for population mean

μ

,

Observation = $\bar{x}$, mean = M and $S.D = \frac{s}{\sqrt{n}}$

Tests of significance can be approximated by simulation.

9.5.3 Test of Significance for single proportion

If X is the number of Successes in 'n' independent trails with Constant probability p of Success for each trail then

E(x) = np and v(x)=npq where q=1-p, is the probability failure

Note that It has been proved that for large sample 'n', The binomial distribution tend to normal distribution Hence for large 'n' ,

X ~N (np,npq)

That is

$$Z = \frac{X - E(X)}{\sqrt{V(X)}} = \frac{X - np}{\sqrt{npq}} \sim N(0,1)$$

and we can apply the normal test.

9.5.4 Remarks (1) In a sample of size n, Let X be the number of persons possessing the given attribute. Then

$$\text{p} = \frac{x}{n}(say)$$

$$\text{E(p)} = E\left(\frac{x}{n}\right) = \frac{1}{n}E(x) = \frac{1}{n} \times np = p$$

$$2)\,\text{V}(p) = \text{V}\left(\frac{x}{n}\right) = \frac{1}{n^2}V(x) = \frac{1}{n^2} \times npq = \frac{pq}{n}$$

$$\Rightarrow S.E(p) = \sqrt{\frac{pq}{n}}$$

3)Since the probable limits for a normal variate X, are $E(X) \pm 3\sqrt{V(X)}$,

The probable limits for the observed proportion of success are $E(\text{p}) \pm 3\, S.E(\text{p})$

$$i.e \quad \text{p} \pm 3\sqrt{pq/n}$$

In particular:

95% Confidence limits for P are given by

$$P \pm 1.96\sqrt{pq/n}$$

99% confidence limits for P are given by

$$P \pm 2.58\sqrt{\frac{pq}{n}}.$$

9.5.4 Problem: A random sample of 500 apples was taken from a large consignment and 60 were found to be bad. Obtain 98% confidence limits for the percentage of bad apples in the consignment.

Solution: We have

P = proportion of bad apples in the sample=60/500=0.12

Since Significant value of 2 at 98% confidence Coefficient is 2.33 (from normd tables),

98% confidence limits for population proportion are

$$P \pm 2.33\sqrt{\frac{pq}{n}} = 0.12 \pm 2.33\sqrt{\frac{0.12 \times 0.88}{500}}$$

$$= 0.12 \pm (2.33 \times \sqrt{0.0002112})$$
$$= 0.12 \pm (2.33 \times 0.01453)$$
$$= (0.08615, 0.15385).$$

Hence 98% confidence limits for percentage of bad apples in the consignment are (8.61,15.38)

9.6 LEVEL OF SIGNIFICANCE:

So far we have discussed the hypothesis, types of hypothesis, Critical region and types of errors etc. In this section we shall discuss very useful Concept “Level of significance", Which play an important role in decision making while testing of hypothesis.

9.6.1 Definition: The probability of type-I error is known as level of significance of a test.

It is also called size of the test (or) Size of critical region, denoted by α.

Note that generally it is pre-fixed as 5% or 1% level ($\alpha = 0.05 \; or \; 0.01$).

Another point about the level of Significance relates to the trueness of the conclusion.

If H_0 do not reject at level say $\alpha = 0.05$ (5% *level*),

Then a person will be confident that “concluding statement about H_0” is true with 95% level.

But even then it may false with 5% chance. There is no cent-percent assurance about the trueness of statement made for H_0.

9.6.2 Example: If among 100 scientists, each draws a random Sample and use the same test statistics to test the same hypothesis H_0 conducting Same experiment, then 95% of them will reach to the same conclusion about H_0. But still 5 of them may differ (against previous conclusion).

9.7 Test for Single mean

Suppose we want to test

(a)If a random Sample x_i (i=1,2,3…n) of size 'n' be drawn from a normal population with specified mean say μ_i;

(b) if the sample mean differs significantly from the hypothetical value μ of the population mean.

Then we have a test for null hypothesis

$H_0 : \mu = \mu_0$ *against the alternative hypothesis*

(1) $H_1 : \mu < \mu_0$

(2) $\mu_X > \mu_0$

(3) $\mu \neq \mu_0$

Here we can easily use the student 't' test.

$$\frac{\bar{x} - \mu}{s/\sqrt{n}} t_{\alpha, n-1}$$

That is the statistics $\frac{\bar{x} - \mu}{s/\sqrt{n}}$ *is a student t with* $(n-1)$ *of freedom, where*

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

is an unbiased estimator of σ^2.

9.7.1 Example: Let p be the probability that a coin will turn up heads and 'x' be the number of heads obtained in tossing the alternative hypothesis.

(a) H_1:p<0.5 at α=0.05 with x=45

(b)H_1:p≠0.5 at α=0.05 with x=38

(c)H_1:p>0.5 at α=0.05 with x=67

Solution:

(a) H_0:p=0.5 , H_1:p<0.5 with ∝=0.05 where p is a population proportion. Let n=100

$$Z = \frac{P - p}{S.E(p)}$$

$$p = \frac{x}{n} = \frac{45}{100} = 0.45$$

$$S.E(p) = \sqrt{\frac{pq}{n}} = \sqrt{\frac{.5 \times .5}{100}} = 0.05$$

$$Z = \frac{0.45 - 0.5}{\sqrt{\frac{pq}{n}}} = \frac{-0.05}{0.05} = -1$$

Since Z(calculated)<Z_0(tabulated)

i.e $Z_{0.05}$=-1.645(from normal tables)

Hence H_0: p=0.05 may be accepted.

(b) $Z = \frac{P - p}{S.E(p)} = Z = \frac{0.38 - 0.50}{\sqrt{\frac{0.5 \times 0.5}{100}}} = -2.4$

|Z|>1.96 for 9.5% level of significance, Hence H_0 is rejected.

This is two tail test.

(c) In this case Z=3.4 >1.645

Hence H_0 is rejected.

9.8 **SOLVED PROBLEMS**

Problem-1

Let us take a situation where a patient suffering from high fever reaches to a doctor. And suppose the doctor formulates the null and alternative hypotheses as

H_0 :The patient is a malaria patient

H_1: The patient is not a malaria patient

Then following cases arise:

Case I: Suppose that the hypothesis H_0 is really true, that is, patient actually a malaria patient and after observation, pathological and clinical examination, the doctor rejects H_0 that is,

he/she declares him/her a non-malaria-patient. It is not a correct decision and he/she commits an error in decision known as type-I error.

Case II: Suppose that the hypothesis H_0 is actually false, that is, patient actually a non-malaria patient and after observation, the doctor does not reject H_0, that is, he/she declares him/her a non-malaria-patient. It is a correct decision.

Case III: Suppose that the hypothesis H_0 is really true, that is, patient actually a malaria patient and after observation, the doctor does not reject H_0. that is, he/she declares him/her a malaria-patient. It is a correct decision.

Case IV: Suppose that the hypothesis H_0 is actually false, that is, patient actually a non-malaria patient and after observation, the doctor does not reject H_0 that is, he/she declares him/her a malaria-patient. It is not a correct decision and he/she commits an error in decision known as type-II error.

Thus, we formally define type-I and type-II errors as below:

Type-I Error:

The decision relating to rejection of null hypothesis H_0 when it is true is called type-I error. The probability of committing the type-I error is called size of test denoted by α and is given by

α= P [Reject H_0 when H_0 is true] = P [Reject H_0 / H_0 is true]

we reject the null hypothesis it random sample / test statistic falls in rejection region, therefore,

$$\alpha = P[X \in \omega / H_0]$$

Where X=(X_1, X_2 ,...,X_n) a random sample and ω is the rejection region and

1-α = [1-P[Reject H_0/H_0 is true]

=P[Do not reject H_0 / H_0 is true] = P[Correct decision]

The (1-α) is the probability of correct decision and it correlates to the concept of

100(1-α) % confidence interval used in estimation.

Type-II Error:

The decision relating to non-rejection of null hypothesis H_0 when it is false (I.e. H_1 is true) is called type-II error. The probability of committing type-I error is generally denoted by β and is given by

β = P[Do not reject H_0 when H_0 is false]

= P[Do not reject H_0 when H_1 is true]

= P[Do not reject H_0 / H_1 is true]

$= P[X \in \overline{\omega} / H_1]$ where, $\overline{\omega}$ is the non-rejection region.

and

1-β=1-P[Do not reject H_0/H_1 is true]

P[Reject H_0/H_1 is true] = P[Correct decision]

The(1-β) is the probability of correct decision and also known as **"power of the test"**. Since it indicates the ability or power of the test to recognize correctly that the null hypothesis is false, therefore, we wish a test that yields a large power.

We say that a statistical test is ideal if it minimizes the probability of both types of errors and maximizes the probability of correct decision. But for fix sample size, α and β are so interrelated that the decrement in one results into the Increment in other. So minimization of both probabilities of type-I and type-II errors simultaneously for fixed sample size is not possible without increasing sample size. Also both types of errors will be at zero level (i.e. no error in decision) if size of the sample is equal to the population size. But it involves huge cost if population size is large. And it is not possible in all situations such as testing of blood.

Problem 2: It is desired to test a hypothesis H_0 : p = p_0 = 1/2 against the Alternative hypothesis H_1 : p = p_1 = 1/4 on the basis of tossing a coin once where p is the probability of "getting head" in a single toss (trial) and agreeing to reject H_0 otherwise. Find the of α and β.

Solution: In such type of problems, first of all we search for critical region

Here, we have critical region ω ={head}

Therefore, the probability of type-I error can be obtained as

α = P[Reject H_0 when H_0 is true]

$= P[X \in \omega / H_0] = P[\text{Head appears} / H_0]$

$= P[\text{Head } appears]_{p=\frac{1}{2}} = \frac{1}{2}$ $[\therefore H_0$ *is true we take value of parameter p given in* $H_0]$

Also,

β= P[Do not reject H_0 when H_1 is true]

$= P[X \notin \omega / H_1] = P[\text{Trail appears} / H_1]$

$= P\left[Trail\ appears\right]_{p=\frac{1}{4}}$ $\left[\therefore H_1 \text{ is true we take value of parameter } p \text{ given in } H_1\right]$

$= 1 - P\left[Head\ appears\right]_{p=\frac{1}{4}} = 1 - \frac{1}{4} = \frac{3}{4}$

9.9 SUMMARY

On Completion of this lesson, we should be able to understand various types of Sampling, develop a frame work for the null hypothesis, alternative hypothesis, Simple and Composite hypothesis, level of Significance and importance of testing of hypothesis. Also discuss the procedure for testing of hypothesis for large Samples.

9.10 TECHNICAL TERMS

- Null Hypothesis (H_0)
- Alternative Hypothesis (H_1)
- Composite Hypothesis
- Type I Error
- Type II Error
- Level of Significance

9.11 EXERCISES

SHORT:

1. Define the null hypothesis (H_0) and the alternative hypothesis (H_1).
2. Differentiate between a simple hypothesis and a composite hypothesis.
3. What are Type-I and Type-II errors in hypothesis testing?
4. What does the level of significance represent in a test of significance?
5. Briefly describe the test for a single mean.

ESSAY:

1. Twenty people were attached by a disease and only 18 Survived. will you reject the hypothesis that the survived rate, If attached by this disease is 85% in favour of the hypothesis that it is more, at 5% level (use large Sample test) (Ans: Z=0.633)
2. A random sample of size 100 has a S.D of 5, what Can you say about the maximum error with 95% Confidence.(Ans: 0.98)
3. A Sample of 64 Students has a mean weight of 70k. gms. Can this be regarded as a Sample from a population with mean weight 65 kgms, and S. D. 25kgms. (Ans z = 1.673).

9.12 FURTHER READING

1. Introduction to Probability and statistics by J. Susan Milton and J.C. Arnold, 4ed , TMH(2007)
2. Mathematical Statistics by R.K. Goel , Krishna Prakasan Media (P) ltd., Meerut
3. Fundamentals of Mathematical Statistics by S.C. Gupta and V.K. Kapoor, Sultan Chand & sons, New Delhi.

Dr. M. Syam Sundar

LESSON-10

INFERENCE ON THE VARIANCE

OBJECTIVES:

After Completion of this lesson the student should be able to:

- understand about inference on the variance, Inference for a single population variance, hypothesis test on population variance.

STRUCTURE

10.1 Introduction

10.2 Statistical inference for a single population Variance

10.3 Chi-Square (χ^2) distribution

10.4 Applications of chi-square distribution

10.5 Inferences about population variance.

10.6 Test for the Variance of a normal population.

10.7 Test for the equality of Variances of two normal populations:

10.8 Solved problems

10.9 Exercise

10.10 Summary

10.11 Technical Terms.

10.12 Self-Assessment Questions

10.13 Further Readings

10.1 INTRODUCTION

The mean of population is important, but in many cases the Variance of the population is just as Important. In most production processes, quality is measured by how closely the process matches the target (i.e the mean) and by the variability (i.e; the Variance) of the process. In order to Construct Confidence interval or test the hypothesis on variance σ^2 , we use the Sampling distributions follows a χ^2 - distribution with n-1 degrees of freedom.

10.2 STATISTICAL INFERENCE FOR A SINGLE POPULATION VARIANCE

First to Construct the confidence interval, take a random sample of size 'n' from a normally distributed population. Calculate the Sample variance σ^2 . The limits for the Confidence interval with Confidence level C for unknown population variance σ^2 are

$$\text{Lower limit}=\frac{(n-1)X\sigma^2}{\chi_R^2}$$

$$\text{Upper limit}=\frac{(n-1)X\sigma^2}{\chi_L^2}$$

Where χ_L^2 is the χ^2 score so that the area in the left-tail of the χ^2 distribution is $\frac{1-C}{2}$,

χ_R^2 is the χ^2 score so that the right tail of the χ^2 - distribution is $\frac{1-C}{2}$ and

χ^2 - distribution has n-1 degrees of freedom.

10.2.1 Steps to conduct a hypotheses test for a population Variance

Step 1: Write down the null and alternative hypothesis in terms of the population variance σ^2

Step 2: Use the form of the alternative hypothesis to determine if the test is left tailed, right-tailed or both.

Step 3: Collect the Sample information for the test and identify the significance level 'α'.

Step 4: Use the χ^2 distribution to find the p-value (the area in the Corresponding tail) for the test. The χ^2 Score and degrees of freedom are

$$\chi^2=\frac{(n-1)\times s^2}{\sigma^2},(\deg rees\ of\ \ freedom=n-1)$$

step-5: Compare the p-value to the significance level State the outcome of the test:

(a) If p-value $\leq \alpha$, reject H_0 in favor of H_1

(b) if p-value $> \alpha$ do not reject H_0

Step-6: Write down a concluding Sentence Specific to the context of the question.

10.2.2 Example: A statistics instructor at a local college claim that the variance for the final exam scores was 25. After speaking with his classmates, one the Class's best students think that the variance for the final exam Scores is higher than the instructor claims, The Student Challenges the instructor to prove her claim. The instructor takes a sample 30 final exams and

finds the Variance of the scores 28. At the 5% Significance level, test if the variance of the final exam Scores is higher than the instructor claims.

Solution: Given that

Hypothesis:

$$H_0 : \sigma^2 = 25$$
$$H_1 : \sigma^2 > 25$$

p-value. By the given question, we have n=30, s^2 =28 and $\propto$ = 0.05 Because the alternative hypothesis is a > the p-value is the are a in the right tail of the χ^2 -distribution.

To use the Chi square(χ^2) distribution to right tail function, we need to calculate out the χ^2 score and the degrees of freedom

$$\chi^2 = \frac{(n-1)\times s^2}{\sigma^2} = \frac{(30-1)\times 28}{25} = 32.48$$

Degrees of freedom = n-1=30-1=29

Function	$\chi^2 - test$	Answer
Field-1	32.48	0.2992
Field-2	29	-

So the p-value=0.2992.

Conclusion:

Because p-value = 0.2992>0.05= $\propto$,we do not reject the null hypothesis. At the 5% significance level there is not enough evidence to Suggest that the variance of the final exam Scores is higher than 25

10.3 CHI-SQUARE DISTRIBUTION

The Square of a Standard normal Variate is known as a Chi-Square (χ^2) Variate with 1 degree of freedom.

Thus if $X \sim N(\mu, \sigma^2)$, then $Z = \frac{X-\mu}{\sigma} \sim N(0,1)$ *and* $Z^2 = \left(\frac{X-\mu}{\sigma}\right)^2$

is a chi-square variate with 1 degree of freedom.

In general X_i (I=1,2,..,n) are 'n' independent normal variates with means μ_i and Variances $\sigma_i^2, (i = 1, 2, ...n)$

Then $$\chi^2 = \sum_{i=1}^{n} \left(\frac{X_i - \mu_i}{\sigma_i} \right)^2$$

is a chi-square (χ^2) variate with n degrees of freedom.

The chi-square (χ^2) distribution can be derived by using

(i) Method of moment generating function

(ii) Method of induction.

10.3.1 moment generating function of χ^2 – distribution.

Let X~ $\chi^2_{(n)}$ *then*

$$M_X(t) = E(e^{tx}) = \int_0^\infty e^{tx} f(x)\, dx$$

$$= \frac{1}{2^{n/2}\Gamma(\frac{n}{2})} \int_0^\infty e^{tx} e^{-x/2} x^{(n/2)-1}\, dx$$

$$= \frac{1}{2^{n/2}\Gamma(\frac{n}{2})} \int_0^\infty \exp\left[-\left(\frac{1-2t}{2} \right) x \right] x^{(n/2)-1}\, d$$

$$= \frac{1}{2^{n/2}\Gamma(\frac{n}{2})} \frac{\Gamma(\frac{n}{2})}{\left[\left(\frac{1-2t}{2} \right) \right]^{n/2}} \qquad \text{(Using gamma integral)}$$

$$= (1-2t)^{-\frac{n}{2}}, |2t| < 1$$

which is the required moment generating function of a χ^2 -variate with 'n' degrees of freedom.

10.3.2 Conditions for the validity of χ^2 -test

χ^2 -test is an approximate test for large values of n. For the validity of chi-square test of 'goodness of fit' between theory and experiment, the following conditions must be satisfied.

a) The sample observations should be independent

b) Constraints on the cell frequencies, if any should be linear, for example,

$$\sum n_i = \sum \lambda_i \quad (or) \sum O_i = \sum E_i, \; I = 1, 2 \ldots n$$

c) N, The total frequency should be reasonably large, say greater than 50.

d) It may noted that the χ^2 -test depends only on the set of observed and expected frequencies and on degrees of freedom.

10.4 APPLICATIONS OF χ^2 -DISTRIBUTION:

χ^2 -distribution has a of large number of applications, In statistics, some of which are given below:

(i) To test if the hypothetical value of the population Variance is $\sigma^2 = \sigma_0^2 \; (say)$

(ii) To test the goodness of fit

(iii)To test the independence of attributes

(iv) To test the homogeneity of independent estimates of the population variance

(v) To combine various probabilities obtained from independent experiments to give a single test of Significance

(vi) To test the homogeneity of independent estimates of the population correlation Coefficient.

10.5 INFERENCE ABOUT A POPULATION VARIANCE

Suppose we want to test if a random sample x_i(i=1,2,…n) has been drawn from a normal population with a specified variance

$$\sigma^2 = \sigma_0^2 \; (say)$$

Under the null hypothesis that the population variance $\sigma^2 = \sigma_0^2$, *the statistic.*

$$\chi^2 = \sum_{i=1}^{n}\left[\frac{x_i - \overline{x_i}}{\sigma_0^2}\right]$$

$$= \frac{1}{\sigma_0^2}\left[\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum(x_i)\right)^2}{n}\right]$$

$$= \frac{ns^2}{\sigma_0^2}$$

Follows χ^2 -distribution with (n-1) degrees of freedom.

By comparing the calculated value with the tabulated value of χ^2 -for (n-1) d.f at certain level of significance (Usually 5%), we may retain or reject the null hypothesis

10.5.1 Example: Test the hypothesis that σ=10, given that s=15 for a random sample of size 50 from a normal population.

Solution: Null hypothesis

$H_0 : \sigma=0$

Given that n=50, s=15

Now by known formula

$$\chi^2 = \frac{ns^2}{\sigma^2}$$

$$\chi^2 = \frac{ns^2}{\sigma^2} = \frac{50\times 15^2}{100} = \frac{50\times 225}{100} = 112.5$$

Since 'n' is large, the statistic of test is,

$$Z = \sqrt{2\chi^2} - \sqrt{2n-1} \sim N(0,1)$$

$$\therefore Z = \sqrt{225} - \sqrt{99} = 15 - 9.95 = 5.05$$

Since |Z|>3, it is significant at all levels of significance and hence H_0 is rejected and we conclude that $\sigma \neq 10.$

10.6 TEST FOR THE VARIANCE OF A NORMAL POPULATION

Let us now consider the problem of testing if the variance of a normal population has a specified value σ_0^2 on the basis of a random sample $x_1, x_2, \ldots x_n$ of size n from a normal population $N(\mu, \sigma^2)$

Now we want to test the hypothesis

$$H_0 : \sigma^2 = \sigma_0^2$$

Against the alternative hypothesis

$$H_1 : \sigma^2 \neq \sigma_0^2$$

(i)If we want to test $H_0 : \sigma^2 = \sigma_0^2$ against the alternative hypothesis $H_1 : \sigma^2 < \sigma_0^2$, we get a one-tailed (left-tail) test with critical region $\chi^2 < \chi^2_{(n-1)}$ while for testing H_0 against $H_1 : \sigma^2 > \sigma_0^2$, we have a right tail test with critical region $\chi^2 > \chi^2_{(n-1)}(\alpha)$

(ii) If we want to test the null hypothesis $H_0 : \sigma^2 = \sigma_0^2$ against the various alternative hypothesis, say $\sigma_0^2 > \sigma_0^2$ (or) $\sigma^2 > \sigma_0^2$, (or) $\sigma^2 < \sigma_0^2$ $(or) \sigma^2 \neq \sigma_0^2$ for the normal population $N(\mu, \sigma^2)$ where 'μ' is known then the the test statistic, the critical region and the confidence interval σ^2 can be obtained from the known table given below on replacing (n-1) by 'n' and ns^2 by $\sum_{i=1}^{n}(x_i - \mu)^2$

Table

S.no	Alternative Hypothesis	Test	Test Statistic	Reject H_0 at '$\propto$' level of significance if	Confidence ($1-\alpha$) interval for σ^2
1	$\sigma^2 > \sigma_0^2$	Right-tailed test	$\chi^2 = \frac{ns^2}{\sigma_0^2}$	$\chi^2 > \chi^2_{(n-1)}(\alpha)$	$\sigma^2 \geq \frac{ns^2}{\chi^2_{(n-1)}(\alpha)}$

2	$\sigma^2 < \sigma_0^2$	Left-tailed test	$\chi^2 = \frac{ns^2}{\sigma_0^2}$	$\chi^2 < \chi^2_{(n-1)}(1-\alpha)$	$\sigma^2 \le \frac{ns^2}{\chi^2_{(n-1)}(1-\alpha)}$
3	$\sigma^2 \neq \sigma_0^2$	Two-tailed test	$\chi^2 = \frac{ns^2}{\sigma_0^2}$	$\chi^2 > \chi^2_{(n-1)}(\frac{\alpha}{2})$ and $\chi^2 > \chi^2_{(n-1)}(\frac{\alpha}{2})$ and $\chi^2 > \chi^2_{(n-1)}(1-\frac{\alpha}{2})$	$\frac{ns^2}{\chi^2_{(n-1)}(\frac{\alpha}{2})} \le \sigma^2 \le \frac{ns^2}{\chi^2_{(n-1)}(1-\frac{\alpha}{2})}$

Normal Population $N(\mu,\sigma^2)$, μ *unknown*,

$H_0 : \sigma^2 = \sigma_0^2$

10.6.1 Remarks

1) Sample observations are independent

2) The Variable under study is Continuous

3) Probability density function is Continuous

(4) Lower order moments exists.

10.7 TEST FOR THE EQUALITY OF VARIANCES OF TWO NORMAL POPULATIONS

Consider two normal population $N(\mu_1,\sigma_1^2)$ and $N(\mu_2,\sigma_2^2)$ where means μ_1 and μ_2 and Variances σ_1^2, σ_2^2 are unspecified.

We want to test the hypotheses

$H_0 : \sigma_1^2 = \sigma_2^2 = \sigma^2$ (unspecified)

With μ_1 and $\mu_2(unspecified)$ against the alternative hypothesis $H_1 : \sigma_1^2 \neq \sigma_2^2$ and μ_1 , μ_2 are unspecified.

If $x_{1i,}$ i=1,2,…n, x_{2j} (j=1,2,…n) be independent random samples of sizes m and n from $N(\mu_1, \sigma^2)$ and $N(\mu_2, \sigma^2)$ respectively , Then

$$L = \left[\frac{1}{2\pi\sigma_1^2}\right]^{\frac{m}{2}} \exp\left[-\frac{1}{2\sigma_1^2}\sum_{i=1}^{m}(x_{1i}-\mu_1)^2\right] X \left[\frac{1}{2\pi\sigma_2^2}\right]^{\frac{n}{2}} \left\{-\frac{1}{2\sigma_2^2}\sum_{j=1}^{n}(x_{2j}-\mu_2)^2\right\} \quad \rightarrow (1)$$

In this case $\theta = [\mu_{1,}\mu_{2,}\sigma_1^2, \sigma_2^2 : \quad -\infty < \mu_i < \infty, \quad \sigma_i^2 > 0, \ i = 1,2,...]$

and $\theta_0 = \left\{(\mu_{1,}\mu_{2,}\sigma^2) : \quad -\infty < \mu < \infty, \ (i = 1,2), \sigma^2 < 0\right\}$

By known formula

$$L(\hat{\theta}) = \left[\frac{1}{2\pi s_1^2}\right]^{\frac{m}{2}} \left[\frac{1}{2\pi s_2^2}\right]^{\frac{n}{2}} .\exp\left(\frac{-1}{2}(m+n)\right) \quad \rightarrow (2)$$

Where s_1^2 *and* s_2^2 *are*

In θ_0, *the likelyhood function is given by*

$$L(\theta_0) = \left[\frac{1}{2\pi\sigma^2}\right]^{\frac{m+n}{2}} .\exp\left[-\frac{1}{2\sigma^2}\left\{\sum_i (x_{1i}-\mu_1)^2 + \sum(x_{2j}-\mu_2)^2\right\}\right] \quad \rightarrow (3)$$

And max*imum likelyhood events for* $\mu_{1,}\mu_2$ *and* σ^2 *are given as* $\hat{\mu}_1 = \overline{x_1}$, $\hat{\mu}_2 = \overline{x_2}$ $\quad \rightarrow (4)$

and $\hat{\sigma}^2 = \frac{1}{m+n}\left\{\sum(x_{1i}-\hat{\mu}_1)^2 + \sum(x_{2j}-\hat{\mu}_2)^2\right\}$

$$= \frac{1}{m+n}\left\{\sum(x_{1i}-\bar{x}_1)^2 + \sum(x_{2j}-\bar{x}_2)^2\right\}$$

$$= \frac{ms_1^{\,2} + ns_2^{\,2}}{m+n} \quad \rightarrow (5)$$

Substituting (3) and(4) in (5), we get

$$L(\hat{\theta}_0)=\left[\frac{m+n}{2\pi(ms_1^2+ns_2^2)}\right]^{\frac{m+n}{2}}.\exp\left[-\frac{1}{2}(m+n)\right] \qquad \rightarrow (6)$$

$$\therefore \lambda=\frac{L(\hat{\theta})_0}{L(\hat{\theta})}$$

$$=\frac{(m+n)^{\frac{m+n}{2}}}{m^{\frac{m}{2}}.n^{\frac{n}{2}}}=\left\{\frac{(ms_1^2)^{\frac{m}{2}}(ns_2^2)^{\frac{n}{2}}}{(ms_1^2+ns_2^2)^{\frac{m+n}{2}}}\right\} \qquad \rightarrow (7)$$

Follows F-distribution with (m-1,n-1) d.f

$$\Rightarrow F=\frac{m(n-1)s_1^2}{n(m-1)s_2^2}$$

$$\Rightarrow \frac{m-1}{n-1}F=\frac{ms_1^2}{ns_2^2}$$

$$\Rightarrow \lambda=\frac{(m+n)^{\frac{m+n}{2}}}{m^{\frac{m}{2}}.n^{\frac{n}{2}}}\left[\frac{\left(\left(\frac{m-1}{n-1}\right)F\right)^{\frac{m}{2}}}{\left(1+\frac{m-1}{n-1}F\right)^{\frac{m+n}{2}}}\right] \qquad \rightarrow (8)$$

10.8 SOLVED PROBLEMS

1) **For a ages Sample of 30 Congenitally blind pupils ages 9 to 15, the S. D. of the WISC Verbal IQ Scores is 16.0, where as for the sighted population σ = 15. Test at α = 0.10 level of Significance** $H_0:\sigma^2=(15)^2=225, H_a:\sigma^2\neq 225$

Solution: The observed test statistic for testing H_0 against H_a is

$$\chi^2_{obs}=\frac{(n-1)s^2}{k}=\frac{(30-1)16^2}{225}=33.0$$

(i) Using the p-value method: Since P(χ^2 >33.0)= 0.72 > 0.5 the p-value is

2p (χ^2 >33.0)=2* .0.28= 0.56.

Thus we fail to reject the null hypothesis.

(ii) Using the critical values method: The Critical Values from table D are

$_{0.05}\chi^2_{29} = 17.71$ *and*

$_{0.95}\chi^2_{29} = 42.56$

Clearly $_{0.05}\chi^2_{29} = 17.71 < \chi^2_{obs} = 33 < {}_{0.95}\chi^2_{29} = 42.56$

So, Here also we fail to reject the null hypothesis

(iii) The 0.90 Confidence interval for σ^2 is

$$\frac{29\times 16^2}{0.95\chi^2_{29}} < \sigma^2 < \frac{29\times 16^2}{0.05\chi^2_{29}}$$

$$\Rightarrow \frac{29\times 16^2}{42.56} < \sigma^2 < \frac{29\times 16^2}{17.71}$$

$$\Rightarrow 174 < \sigma^2 < 419$$

With individual lines at its Various windows, a post office finds that the standard deviation for normally distributed waiting times for consumers is 7.2 minutes. The post office experiments with a Single, main waiting line and finds that for a random sample of 25 Customers the waiting times for customers have a Standard deviation of 4.5($\sigma^{2)}$ minutes. At 5%. Significance level. Determine if the single line charged the variation among the wait times for Customers.

Solution: From the given data, Hypothesis:

H_0: $\sigma^2 = 51.84$

H_0: $\sigma^2 \neq 51.84$

P-value from the question, we have n=25,

s^2 =20.25 and α= 0.05

Now we need to calculate out the χ^2 . Score and the degrees of freedom

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2} = \frac{(25-1)\times 20.25}{51.84} = 9.375$$

degrees of freedom = n-1=25-1=24

Because this is two tailed test, we need to know which tail (left or right) we have the

χ^2-score so that we can use the correct excel function, If χ^2>df-2 the χ^2- score Corresponds to the right tail.

If the χ^2 < df-2, the χ^2-score Corresponds to the right tail.

If χ^2>df-2, the χ^2 - score corresponds to the left tail.

In this case χ^2 =9.375 < 22= d.f-2

So, χ^2- score Corresponds to the left tail.

Here we need to use χ^2-distribution on to find the area in the left tail.

Function	**Chi-square distribution**	**Answer**
Field-1	9.375	0.0033
Field-2	24	-

So the area in the left tail is 0.0033, which means that 1/2(p-value)=0.0033.

This is also the area in the right tail. So

p-value=0.0033+0.0033=0.0066

Conclusion

Because p-value = 0.0066 <0.05 = α, we reject the null hypothesis in favour of the alternative hypothesis.

At 5% Significance level there is enough evidence to suggest that the variation among the wail to times for Customer has Changed.

10.9 EXCERSISE

1) If X is a chi-square Variate with 'n' d.f, then prove that for large ‘n’

$$\sqrt{2X} \sim N(\sqrt{2n},1)$$

2) Show that for 2 d.f the probability p of a value of χ^2 greater than χ_0^2 is $\exp(\frac{-1}{2}\chi_0^2)$ and hence that $\chi_0^2 = 2\log_e(1/p)$.

Deduce the value of χ_0^2 when p=0.05

(3) Find the size of the sample if S.D. of the population is 9 and there should be 99% Confidence that the error of estimate will not exceed 3?

(4) In two independent Samples of Sizes 8 and 10 the sum of squares of deviations of the sample Values from the respective sample means were 84.4 and 102.6. Test whether the difference of variances of the populations is significant or not.

10.10 SUMMARY

Statistical inference for a single population variance relies on the Chi-Square ($\chi 2$) distribution, used to construct confidence intervals and perform hypothesis testing on variance. The Chi-Square test assesses whether a sample variance significantly differs from a hypothesized value, while the **F-test** compares variances between two normal populations. These methods are fundamental in quality control, scientific research, and hypothesis testing across various disciplines.

10.11 TECHNICAL TERMS.

- Chi-Square (χ) Distribution
- Degrees of Freedom (df)
- Variance ($\sigma 2$)
- Standard Deviation (σ
- F-Test
- Hypothesis Testing
- Critical Value

10.12 SELF-ASSESSMENT QUESTIONS

SHORT:

- Define the Chi-Square ($\chi 2$) distribution and its key properties.
- What is the formula for testing a population variance using the Chi-Square test?
- How do degrees of freedom affect the Chi-Square distribution?
- What is the purpose of the F-test in comparing variances?
- When do we use a Chi-Square test for variance inference?

ESSAY:

- Explain the Chi-Square ($\chi 2$) distribution and its applications in statistical inference.
- Describe the process of hypothesis testing for a single population variance using the Chi-Square test.
- Discuss the importance of variance estimation and its role in statistical analysis.
- Compare and contrast the Chi-Square test for a single variance and the F-test for comparing two variances.
- How does sample size impact the accuracy of variance estimation and hypothesis testing?

10.13 FURTHER READINGS

1. "A First Course in Probability" – Sheldon Ross
2. "Introduction to Probability" – Dimitri P. Bertsekas & John N. Tsitsiklis
3. "Probability and Statistics" – Morris H. DeGroot & Mark J. Schervish
4. "Probability and Random Processes" – Geoffrey Grimmett & David Stirzaker
5. "Probability: Theory and Examples" – Rick Durrett

Dr. M. Syam Sundar

Lesson - 11

INFERENCE ON ONE PROPORTION

OBJECTIVES:

This lesson is prepared in such a way that after studying the material the student is expected to have a thorough comprehension of the above concepts like sampling simple random sampling, Estimation of proportions, testing procedure of Inference on one proportion which are the important areas of investigation and statistical data analysis. The student will be having and well equipped with both theoretical and practical aspects of Inference on one proportion.

STRUCTURE OF THE LESSON:

11.1 Introduction

11.2 Estimation of Proportions

11.3 confidence interval for P

11.4 Maximum error

11.5 sample Size

11.6 Inference on one Proportion - Testing procedure

11.7 Worked out Examples

11.8 Exercise

11.9 Summary

11.10 Technical Terms.

11.11 Self Assessment Questions

11.12 Further Readings

11.1 INTRODUCTION

The word population is used to refer to any collection of objects or results of operations. For example, we may speak of the population of dairy cows in Meerut district, the population of mileages of automobiles types. The population of prices of a commodity in a city. We may also speak of the hypothetical population of heads and tails obtained by tossing a coin an infinite number of times or the population of all possible values which the bank rate can have in twenty years' time and so on.

The aim of the statistical enquiry is to find out something about a specified population. It is impossible or impracticable to examine each member of the population since such a Process will be too costly in terms of time and money. Thus,

the investigator is led to the study of a selected number of individuals from the population and based on this Limited investigation, he makes inferences regarding the whole population. This selected number of individuals from a population is called a sample. The inferences that can be made from a sample about the whole population can never be of a categorical certainty. They can only be expressed in terms of probabilities. In order that the theory of probability can be applied, the sampling should be random. In the Case of non-random sample, there is no way to measure the degree of confidence to be placed in any inference which can be made from such a sample. Remember that the selection of an individual from a population is random when each member of the population has the same chance of being selected. The aims of theory of sampling are (1) to find estimates of certain constants such as mean and standard deviation of the population and (2) to determine what degree of confidence can be placed in these estimates when they are obtained, in other words, to determine the limits within which the parameters of the population are expected to lie with a specified degree of confidence.

$\therefore$ Simple sampling we mean random sampling in which each event has the same chance p of success and in which the chances of success of different events are independent of whether previous Trials have been made or not. It is also to be noted that random sampling is not necessarily simple but simple sampling is always random. As a matter of fact, simple sampling is a particular form of random sampling. For example, is a bag contains 5 black balls and 3 white balls, the chance of drawing a black ball at the second trail is 5/8 and if the ball is not replaced the chance of drawing a black ball at the second trial is 4/7 which is not same as before. Hence the sampling is not simple. However, the Sampling is random since on the first trail each black ball has got the same chance 5/8 of being drawn out and at the second trail each black ball has the same chance 4/7 of being selected.

Suppose we draw samples of size N from a large population. If each individual in a sample has a chance for success. i.e., selection and a chance q=1-p for failure so that the probabilities of 0,1, 2,…n successes are given by successive terms of $(q+p)^n$ and $p+q=1$

This means that the probabilities of 0,1,2,..n times in the sample possessing the attribute under study is $q^n, {}^nc_1 q^{n-1}.p, {}^nc_2 q^{n-2}.p^2 ... p^n$ we know that the mean of this distribution is np and standard deviation is $\sqrt{npq}$. Hence the expected value of success in a sample of size n is np and the standard error of the member of the successes in sample of size is $\sqrt{npq}$.

If instead of the number of successes in each sample we take proportion of successes, the mean proportion of successes will be $\frac{np}{n} = p$ and the standard error of the proportion of successes is $\sqrt{\text{n}.\frac{p}{n}.\frac{q}{n}} = \sqrt{\frac{pq}{n}}$

Note: If p or q becomes very small, then pq=p(1-p)=p approx , Hence, $\sigma = \sqrt{np} = \sqrt{M}$ It follows that if the proportion of successes be small, the standard error of the number of successes is the square root of the mean number of successes.

Test of Significance for large samples:

We know that if n is large, the binomial distribution tends to normal so that in the case of large samples properties of normal curves can be used. Suppose we wish to test the hypothesis that a given large sample of size n is obtained by simple sampling from a population for which the probability of success is p. For normal distribution, we know that 99.7%. of its members lie within a range ± 3σ i.e ± $3\sqrt{npq}$ on either side of the mean np, so that only 0.3% of the members lie outside this range.

Again only 5% of the members of a normal population lie outside the range mean ± 2σ [i.e, np ± $2\sqrt{npq}$] . Hence, we have the following test of significance for large samples.

If the number of successes in a large sample of size differs from the expected value np by more than $3\sqrt{npq}$. We Call this difference highly significant. Sometimes a difference of more than 2 $\sqrt{npq}$ is called insignificant. This test is due to the central limit theorem.

If z is the standard normal variate, we have

$$Z = \frac{x - np}{\sqrt{npq}}$$

Where x is the observed number of successes in the sample.

Thus

(i) If |Z|>3, the difference between the observed and the expected number of successes is highly significant.

(ii) If 2 < |Z| < 3 the differences may be regarded as significant

(iii)If |Z| < 2 , the difference is not significant.

11.2 ESTIMATION OF PROPORTIONS

If x is the number that an event occurs among n trials. The proportion of the time that the event occurs i.e, $\frac{x}{n}$ is the sample proportion. If n trials satisfy the

assumption underlying the binomial distribution. We know that the mean and standard deviation of the number of successes are given by np and $\sqrt{np(1-p)}$ where p is the probability of success.

$i.e; E(\mathrm{x}) = np, V(\mathrm{x}) = np(1-p)$ [$\therefore$ *here* $1-p$ *is the probability of failure*)

$\therefore S.D \quad \sigma = \sqrt{np(1-p)}$

If we divide these by E(x) and σ by n we get

$$E(\frac{\mathrm{x}}{n}) = \frac{1}{n}E(X) = \frac{1}{n}.np = p$$

$$\frac{\sigma}{n} = \frac{\sqrt{np(1-p)}}{n} = \sqrt{\frac{p(1-p)}{n}}$$

Sample proportion is unbiased estimator of the binomial parameter P. If we write P for probability of Success and & Q for probability of failure. i.e, Q = 1 - P

p= sample proposition

$$\mathrm{E(p)} = E(\frac{\mathrm{x}}{n}) = \mathrm{P}$$

$$\mathrm{V(p)} = \mathrm{V}(\frac{\mathrm{x}}{n}) = \frac{1}{n^2}V(\mathrm{x}) = \frac{nPQ}{n^2} = \frac{pQ}{n}$$

Standard Error of $p = \sqrt{\frac{PQ}{n}}$

If the sample is taken from a finite population of size N then standard error of proportions is

$$S.E(p) = \sqrt{\frac{N-n}{N-1}\ \frac{PQ}{n}}$$

11.3 CONFIDENCE INTERVAL FOR P:

Approximately having the degree of confidence $(1-\alpha)\,100\%$, assuming x_0 as the largest integer for which the binomial probabilities b (k;n,p) satisfying $\sum_{k=0}^{x_0} b(k,n,p) \le \alpha/2$ while x_1, is the smallest integer for which $\sum_{k=x_1}^{n} b(k,n,p) \le \alpha/2$

Then x_0 *and* x_1 depend upon the value of P we can write $x_0(\mathrm{P})$, $x_1(\mathrm{P})$ we can assert with a probability of approximately $1-\alpha$ and *at least* $1-\alpha$

$\Rightarrow \quad x_0(\mathrm{P}) < x < x_1(\mathrm{P})$ will be satisfied.

If 'n' is large we know that the binomial distribution tends to normal distribution. Then for large n

$$Z = \frac{X - E(\mathrm{X})}{\sqrt{V(\mathrm{X})}} = \frac{X - np}{\sqrt{nPQ}} \sim N(0,1)$$

Confidence limits for p in terms of the observed value x substituting $\frac{X}{n}$ for p

$$\frac{X}{n} - Z_{\alpha/2}\sqrt{\frac{\frac{X}{n}(1-\frac{X}{n})}{n}} < p < \frac{X}{n} + Z_{\alpha/2}\sqrt{\frac{\frac{X}{n}(1-\frac{X}{n})}{n}}$$

is the confidence interval for p (proportions)

If we write $\frac{X}{n} = P$

The confidence interval for large sample for p

$$P - Z_{\alpha/2}\sqrt{\frac{PQ}{n}} < p < P + Z_{\alpha/2}\sqrt{\frac{PQ}{n}} \qquad (\therefore P = \frac{X}{n}, Q = 1 - P)$$

11.4 MAXIMUM ERROR:

The magnitude of the error, when we use $\frac{X}{n}$ as an estimator of p is given by

$$|\frac{X}{n} - P|$$

Using the approximation, which can be associated with probability 1- α.

$\therefore$ The inequality $|\frac{X}{n} - P| \le Z_{\alpha/2}\sqrt{\frac{PQ}{n}}$ Will be satisfied.

$\therefore$ The error will be at the most $Z_{\alpha/2}\sqrt{\frac{PQ}{n}}$

Hence the maximum error of estimate for the proportion p is

$$E = Z_{\alpha/2}\sqrt{\frac{PQ}{n}}$$

11.5 SAMPLE SIZE

Maximum error of estimate for the proportion p is

$$E = Z_{\alpha/2}\sqrt{\frac{PQ}{n}}$$

$$\Rightarrow \sqrt{n} = Z_{\alpha/2}\frac{\sqrt{PQ}}{E}$$

$$\Rightarrow N = \left[\frac{Z_{\alpha/2}}{E}\right]^2 .PQ$$

$$hence\ the\ sample\ size\ n = (PQ)\left(\frac{Z_{\alpha/2}}{E}\right)^2$$

$$or\ N = \left[\frac{Z_{\alpha/2}}{E}\right]^2 .P[1-P]$$

Note: If P is not given, we cannot use this formula, If P is not given we can make use of the fact that P(1- P) is at most 1/4 (i.e P = 1/2)

Sample size 'n' when P is not given $= \frac{1}{4}\left[\frac{Z_{\alpha/2}}{E}\right]^2$

11.6 INFERENCE ON ONE PROPORTION: TESTING PROCEDURE

We shall test the null hypotheses $P= P_0$ against one of the alternatives $P<P_0$ or $P>P_0$ or $P\neq P_0$ statistic for large Sample test concerning one proportion p which is a random variable having approximately the standard normal distribution initially we test at α=001 and 0.05 level of significance. Thus, the testing procedure for one Proportion is as given below.

Step 1: The null hypotheses is H_0: $P=P_0$ i.e; true proportion and specified proportion are equal.

Step 2: The alternative hypothesis is H_1: $P\neq P_0$ which means there is significant difference between the two proportions

Step 3: The test-statistic to test the null hypothesis H_0 against the alternative hypothesis H_1, for large 'n' is

$$Z = \frac{p-P}{\sqrt{\frac{PQ}{n}}} \sim N(0,1)$$

where p- sample proportion

P- Probability of success

Q - Probability of failure = 1-P

n-The sample size

Step4: The $Z_{\text{critical or table value}}$ One taken from standard normal tables i.e, $Z_{\alpha/2}$ usually at $\alpha = 1\%$ and 5% level of significance.

Step 5: Since $Z_{cal.val} \leq Z_{tabulated\ value}$ we accept H_0 at α % l.o.s i.e, the two proportions are equal

If $Z_{cal.val} > Z_{tabulated\ value}$ we reject H_0 at α % L.o.S. i.e, the two proportions are not equal

11.7 WORKEDOUT EXAMPLES.

Example 1: In some dice-throwing experiment weldon threw dice 42, 152 times, and of these 25, 145 yielded 4 or 5 or 6. Is this consistent with the hypothesis that the dice were unbiased?

Solution:

Step 1: The null hypotheses is H_0 : the dice were unbiased and are consistent with the hypothesis.

Step 2: The alternative hypothesis is H_1: the dice were biased and are not consistent with the hypotheses

Step 3: The test statistic to test the null hypothesis against the alternative hypotheses is

$$Z = \frac{X - np}{\sqrt{npq}} \sim \mathrm{N}(0,1)$$

The calculations are as follows

The total number of throws = 49, 152
The chance of throwing for 4 or 5 or 6 with one die =1/2
i.e; 1/6 + 1/6 + 1/6 = 3/6 = 1/2
The expected value of the number of successes = 1/2 ×49152=24576
and the observed value of success=25145
Thus, the excess of the observed value over the expected value= 25145 - 24576 =569.
The standard deviation of simple sampling = $\sqrt{npq} = \sqrt{49152 \times 1/2 \times 1/2} = 110.9$

Hence $Z = \frac{X - np}{\sqrt{npq}} \sim N(0,1) = \frac{569}{110.9} = 5.13$

Step 4: The Z critical value is |Z|=3

Step 5: Conclusion: Since the observed deviation is 5.13 times the standard error , it is therefore highly improbable that it arose as a sampling function. We must therefore seek some other season for this deviation. Hence it seems reasonable to suspect dice were biased.

Example 2: In a large consignment of oranges a random sample of 64 oranges revealed that 14 oranges were bad. Is it reasonable to ensure that 20% of the oranges are bad.

Solution: Step 1: The null hypotheses is H_0 : p = P. Which means the sample and population proportions are equal.
Step 2: The alternative hypothesis is H_1: p ≠ P.Which means the sample and population proportions are not equal.
Step 3: The test statistic to test the null hypothesis against the alternative hypotheses is given by the Z-test statistic for large 'n'
i.e

$$Z = \frac{p - P}{\sqrt{\mathrm{PQ}/n}} \sim \mathrm{N}(0,1)$$

where p is the sample proportion=0.219

$$\text{P is the Population proportion} = \frac{X}{n} = \frac{14}{64} = 0.2$$

Q = 1-P=1-0.2=0.8, X=14, n=64

$$\therefore Z = \frac{0.219 - 0.2}{\sqrt{\frac{0.2 \times 0.8}{64}}} = 3.8$$

Step 4: The Z. critical value is given by |Z|=3.

Step 5:Conclusion: Since, the $Z_{cal\ value}$ =3.8> $Z_{tabulated\ value}$ =3

Hence we reject our null hypothesis H_0 at 5% L.o.S.

Therefore we conclude that there is significant difference between the two proportions.

Example 3: In a study designed to investigate whether Certain detonators used with explosives in coal mining meet the requirement that at least 90% will ignite the explosive when charged. It is found that 174 of 200 detonators-function properly. Test the null hypothesis P = 0.9 against the alternative hypothesis P< 0.9 at 5% L.O.S

Solution: step 1: the null hypothesis is $H_{0:}$ P = 0.9 i.e, the population proportion is equal to 0.9.

Step 2: The alternative hypothesis is H_1 : P < 0.9: i.e the population proportion is less than 0.9.

Step3: The test statistic to test the null hypothesis against the alternative hypothesis is given by Z-test statistic for one proportion for large 'n'

i.e

$$Z = \frac{p - P}{\sqrt{PQ/n}} \sim N(0,1)$$

where p =0.87 The sample proportion

P=0.9 the Population proportion

Q = 1-P=1-0.9=0.1,

n is the sample size=200

Then

$$\therefore Z_{cal.value} = \frac{0.87 - 0.9}{\sqrt{\frac{0.9 \times 0.1}{200}}} = \frac{-0.03}{\sqrt{\frac{0.09}{200}}} = -1.41$$

Step 4: The $Z_{critical\ value\ or\ table\ value}$ is

$Z_\alpha = -1.645$

At 5% L.o.S since it is a left tail test.

Step 5:Conclusion: Since, the $|Z_{cal\ value|}$ =|-1.41|<$|Z|_{tabulated\ value}$ =|-1.645|

we accept our null hypothesis H_0 at 5% L.o.S.

Hence we conclude that there is no evidence significantly to say that the given kind of detonator fails to meet the required standard.

Example 4: A die is thrown 256-times. An even digit turns up 150 times can we say that the die is unbiased

Solution: Step 1: The null hypotheses is H_0 : the die is unbiased.

Step 2: The alternative hypothesis is H_1: the die is biased.
Step3: The test statistic to test the above hypothesis is
i.e

$$Z = \frac{X - np}{\sqrt{nPQ}} \sim N(0,1)$$

where X -Number of success =150

n- the sample size=256

P=Probability of getting an even digit (2or 4 or 6)=1/2

Q=Probability of not getting an even digit (2or 4 or 6)=1/2

Then

$$\therefore Z = \frac{150 - 256 \times 1/2}{\sqrt{256 \times 1/2 \times 1/2}} = \frac{150 - 128}{8} = 2.75$$

Step 4: The $Z_{\text{critical value or table value}}$ is

$$Z_{\alpha/2} = 1.96$$

Obtained from standard normal tables at L.o.S (∵ this is a two tailed test).

Step 5: Since, the $Z_{\text{cal. value}} = 2.75 > Z_{\text{table value}} = -1.96$

Hence we reject our null hypothesis H_0 at $\alpha = 5\%$ L.o.S.

Therefor we conclude that the die is biased.

Example 5: In a sample of 500 people in Maharashtra 300 are wheat eaters. What can you say about the maximum error with 99% confidence.

Solution: Maximum Error estimate of proportion is given by

$$E = Z_{\alpha/2}\sqrt{\frac{PQ}{n}}$$

where X = number of people who eats wheat = 300
n =The sample size=500.
Then proportion of wheat eaters in the sample =P=X/n= 300/500=0.6

Q=1-P=1-0.6=0.4

The Z-table value is $Z_{\alpha/2} = 2.58$ (*from Standard normal tables*)

The Maximum Error estimate is

$$E = Z_{\alpha/2}\sqrt{\frac{PQ}{n}}$$

$$= 2.58\sqrt{\frac{0.6\times 0.4}{500}}$$

$$E = 0.0568$$

Example 6: If we can assert with 95% that the maximum error is 0.05 and P is given as 0.2 Then find the size of the sample.

Solution: Maximum Error Estimate is $E = Z_{\alpha/2}\sqrt{\frac{PQ}{n}}$

Where P = 0.2 , Q= 1 - P = 1-0.2 = 0.8 Error E =0.05

$$\Rightarrow n = \frac{Z^2_{\alpha/2}PQ}{E^2} = \left(\frac{Z_{\alpha/2}}{E}\right)^2 .PQ = \left(\frac{1.96}{0.05}\right)^2 (0.2)(0.8)$$

Then $n = 244$(approximate)

$Z_{\alpha/2}$ is Z-table value obtained from standard normal tables. At α = 5%. L.o.S

$Z_{\alpha/2} = 1.96$ ∵

∴ Size of the sample n=244.

Example 7: In a random sample of 100 packages shipped by air freight 13 had some damage. Construct 95% confidence interval for the true proportions of damage package.

Solution: The Confidence interval for proportion p 100 (1-α)% given by

$$P - Z_{\alpha/2}\sqrt{\frac{PQ}{n}} < p < P + Z_{\alpha/2}\sqrt{\frac{PQ}{n}}$$

$$or\ p \in \left[P - Z_{\alpha/2}\sqrt{\frac{PQ}{n}}, P + Z_{\alpha/2}\sqrt{\frac{PQ}{n}}\right]$$

The number of damaged packages x=13

The sample size n=100

Proportion of damaged packages P=x/n = 13/100=0.13.

Proportion of no bad packages in the Sample Q = 1-P= 1-0.13=0.87

The Z table value at α =95%. L.o.S is $Z_{\alpha/2} = 1.96$

Obtained from standard normal tables.

Then 100 (1-α)%. CI is given by

$$P - Z_{\alpha/2}\sqrt{\frac{PQ}{n}} < p < P + Z_{\alpha/2}\sqrt{\frac{PQ}{n}} = 0.13 - 1.96\sqrt{\frac{0.13\times 0.87}{100}} < p < 0.13 + 1.96\sqrt{\frac{0.13\times 0.87}{100}}$$

Confidence interval is 0 .13 -0. 066 < p < 0 .13 + 0.66

$\Rightarrow$ 64 < p < 0.196

11.8 EXERCISE:

1. A coin is tossed 400 times and it turns up head 216 times. Discuss whether the coin may be an unbiased one.

2. A die is thrown 9,000 times and a throw of 3 or 4 is reckoned as a success. Suppose that 3240 throws of a 3 or 4 have been made out. Do the data indicate an unbiased die?

3. In a Locality of 18,000 families, a sample of 840 families was selected. Of these 840 families, 206 families were found to have a monthly income of Rs.50 or less. It is desired to estimate how many out of the 18000 families have a monthly income of Rs.50 or less, within what limits would you place your estimate?

4. In a random sample of 200 claims filed against an insurance company writing collision insurance on cars 84 exceeds 1200, construct a 95% confidence interval for the true proportion of claims filed against this insurance company that exceed 1200.

5. What can we say with 99% confidence about the maximum error. If we use the sample proportion as an estimate of the time proportion of claims filed against this insurance company that exceeds 1200 in the above problem 4. (Ans:0.09)

6. In a recent study 69 of 120 meteorites were observed to enter the earth's atmosphere with a velocity of less than 26 miles per second. If we estimate the corresponding true proportion as P what can we say with 95% Confidence about the maximum error. (Ans:0.088)

7. Experience had shown that 20% of a manufactured product is of the top quality.In one day's production of 400 articles only 50 are of top quality. Show that either the production of the day taken was not a representative sample or the hypothesis of 20%. was wrong.

11.9 SUMMARY

In this lesson an attempt is made to explain the Concepts of proportion, estimation of one proportion and hypothesis concerning one proportion, Maximum error estimate, confidence Interval procedures associated with them along with both theory and practical. A few examples are worked out and a good number of exercises are also given.

11.10 TECHNICAL TERMS

- Proportions
- Estimating one proportion
- Maximum Error Estimate
- Hypothesis concerning one proportion
- Confidence Intervals

11.11 SELF ASSESSMENT QUESTIONS

SHORT:

1. Define a confidence interval for a population proportion.
2. What is the formula for the margin of error in estimating a proportion?
3. How does increasing the sample size affect the width of a confidence interval?
4. State the null and alternative hypotheses for a one-proportion test.
5. What is the role of the Z-score in proportion hypothesis testing?

ESSAY:

1. Explain the concept of confidence intervals for proportions and their significance in statistical inference.
2. Discuss the factors that influence the margin of error in estimating a proportion and how they can be controlled.
3. Describe the step-by-step procedure for hypothesis testing on a single proportion with an example.
4. How is the required sample size for estimating a proportion determined? Discuss its importance with real-world applications.
5. Compare and contrast confidence intervals and hypothesis testing in the context of proportion estimation.

11.12 FURTHER READINGS

1 "A First Course in Probability" – Sheldon Ross

2 "Introduction to Probability" – Dimitri P. Bertsekas & John N. Tsitsiklis

3 "Probability and Statistics" – Morris H. DeGroot & Mark J. Schervish

4 "Probability and Random Processes" – Geoffrey Grimmett & David Stirzaker

5 "Probability: Theory and Examples" – Rick Durrett

Dr. T.V. Pradeep Kumar

LESSON-12

INFERENCE ON TWO PROPORTIONS

OBJECTIVE:

This lesson is prepared in such a way that after studying the material the student is expected to have a thorough comprehension of the above concept's estimation of two proportions, confidence interval, Maximum error, Sample size which one the important areas of investigation and statistical data analysis. The students will be having and well equipped with both theoretical and practical aspects of Inference on two proportions.

STRUCTURE OF THE LESSON:

12.1 Introduction
12.2. Estimation of two proportions
12.3. Confidence interval for P_1-P_2
12.4 Maximum Emer
12.5 Sample Size
12.6 Inference on two Proportions - Testing Procedure.
12.7 Inference on greater than or equal to two proportions
12.8. Workedout Examples
12. 9. Exercise
12.10. Summary
12.11 Technical Terms.
12.12 Self Assessment Questions
12.13 Further Readings

12.1 INTRODUCTION

In every field, especially, many engineering problems deal with proportions, percentages or probabilities. In acceptance sampling we are concerned with the proportion of defectives in a Lot and in life testing we are concerned with the percentage of certain components which will perform satisfactorily during a stated period of time, or the probability that a given component will last at least a given number of hours. It is clear from these examples that problems concerning proportions, percentages or probabilities are really equivalent; a percentage is merely a proportion multiplied by 100. and a probability may be interpreted as a proportion in a long series of trials

12.2 ESTIMATION OF TWO PROPORTIONS:

Suppose we want to compare two distinct populations with respect to the prevalence of a certain attribute, say A, among their members Let x_1, x_2 be the number of persons possessing the given attribute A in random samples of sizes n1 and n2 taken from two populations respectively. Then sample proportions are given by

$P_1 = \frac{x_1}{n_1}$, $P_2 = \frac{x_2}{n_2}$

If P_1 *and* P_2 *are population proportions then*

$$E(P_1) = E\left(\frac{x_1}{n_1}\right) = \frac{1}{n_1}E(x_1) = \frac{1}{n_1}.n_1P_1 = P_1$$

$$E(P_2) = E\left(\frac{x_2}{n_2}\right) = \frac{1}{n_2}E(x_2) = \frac{1}{n_2}.n_2P_2 = P_2$$

$$\mathrm{V}(P_1) = V\left(\frac{x_1}{n_1}\right) = \frac{1}{n_1^2}\mathrm{V}(x_1) = \frac{1}{n_1^2}.n_1P_1Q_1 = \frac{P_1Q_1}{n_1} \quad [\because \textit{For Binomial distribution mean} = E(x) = np \textit{ and } \mathrm{var}\textit{iance} = V(x) = npq]$$

$$\mathrm{V}(P_2) = V\left(\frac{x_2}{n_2}\right) = \frac{1}{n_2^2}\mathrm{V}(x_2) = \frac{1}{n_2^2}.n_2P_2Q_2 = \frac{P_2Q_2}{n_2}$$

Since for large samples P_1 and P_2 are asymptotically normally distributed and hence $(P_1 - P_2)$ is also normally distributed the standard variable corresponding to the difference $(P_1 - P_2)$ is given by

$$Z = \frac{(P_1 - P_2) - E(P_1 - P_2)}{\sqrt{V(P_1 - P_2)}} \sim N(0,1)$$

Under the null hypotheses $H_0: P_1 = P_2$ i.e, there is no significant difference between the sample proportion.

$$\therefore E(P_1 - P_2) = E(P_1) - E(P_2) = P_1 - P_2 = 0$$

$$V(P_1 - P_2) = V(P_1) + V(P_2) = \frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2} = PQ\left[\frac{1}{n_1} + \frac{1}{n_2}\right]$$

Since under $H_0 : P_1 = P_2 = P, \quad Q_1 = Q_2 = Q.$

$\therefore$ *under* $H_0 : P_1 = P_2$

$$Z = \frac{(P_1 - P_2)}{\sqrt{PQ\left[\frac{1}{n_1} + \frac{1}{n_2}\right]}} \sim N(0,1)$$

and $P = \frac{(n_1P_1 + n_2P_2)}{n_1 + n_2} = \frac{x_1 + x_2}{n_1 + n_2}$

$$= E(P) = \frac{1}{n_1 + n_2}E\left[n_1P_1 + n_2P_2\right] = \frac{1}{n_1 + n_2}\left[n_1E(P_1) + n_2E(P_2)\right]$$

$$= \frac{1}{n_1 + n_2}\left[n_1P_1 + n_2P_2\right] = P \qquad \therefore P_1 = P_2 = P \textit{ under } H_0$$

$\therefore$ *The estimate is unbiased*.

12.3 CONFIDENCE INTERVAL FOR (P_1-P_2):

Large sample confidence interval for the differences of two proportions

$$= (P_1 - P_2) \pm Z_{\alpha/2}\sqrt{PQ\left[\frac{1}{n_1}+\frac{1}{n_2}\right]}$$

Standard error of (P_1-P_2)

$$= \sqrt{PQ\left[\frac{1}{n_1}+\frac{1}{n_2}\right]}$$

12.4 MAXIMUM ERROR:

Maximum error of estimate is $= Z_{\alpha/2}\sqrt{PQ\left[\frac{1}{n_1}+\frac{1}{n_2}\right]}$

Note: Suppose the population proportions P_1 and P_2 are given to be different i.e, $P_1 \neq P_2$ and we want to test whether the difference $P_1 - P_2$ is significant then the test statistic becomes

$$Z = \frac{(P_1 - P_2) - E(P_1 - P_2)}{S.E(P_1 - P_2)} = \frac{(P_1 - P_2) - (P_1 - P_2)}{\sqrt{\frac{P_1Q_1}{n_1}+\frac{P_2Q_2}{n_2}}}$$

Here the sample proportions are not given. If we set up the null hypothesis H_0: $P_1 = P_2$, the difference in population proportions is likely to be hidden in sampling. then the test statistic becomes

$$|Z| = \frac{|P_1 - P_2|}{\sqrt{\frac{P_1Q_1}{n_1}+\frac{P_2Q_2}{n_2}}}$$

$$Confidence\ Interval = P_1 - P_2 \pm Z_{\alpha/2}\sqrt{\frac{P_1Q_1}{n_1}+\frac{P_2Q_2}{n_2}}$$

$$S.E(P_1 - P_2) = \sqrt{\frac{P_1Q_1}{n_1}+\frac{P_2Q_2}{n_2}}$$

$$and\ M\text{inimum}\ error = Z_{\alpha/2}\sqrt{\frac{P_1Q_1}{n_1}+\frac{P_2Q_2}{n_2}}$$

12.5 SAMPLE SIZE:

If $n_1=n_2=n$, the level of significance α, Type II error probability β at the alternative values P_1 ,P_2 with $P_1 - P_2=d$ when

$$n=\frac{\left[Z_\alpha\sqrt{(P_1+P_2)(Q_1+Q_2)/2}+Z_\beta\sqrt{P_1 Q1+P_2 Q_2}\right]^2}{d^2}$$

For an upper or lower-tailed test, with $\alpha/2$ replacing α for a two tailed test.

12.6. INFERENCE ON TWO PROPORTIONS - TESTING PROCEDURE:

The Testing procedure for inference on Two proportions is as follows.

Step 1: The Null hypothesis is $H_0: P_1-P_2 = 0$ i.e, there is no significant difference between two proportions.

Step 2: The alternative hypothesis is $H_1: P_1-P_2 \neq 0$ i.e, there is no significant difference between two proportions.

Step3: The test statistic to test the above hypothesis is given by Z-test statistic for 2 proportions for large n.

i.e $$Z-test\ statistic=\frac{(P_1-P_2)}{\sqrt{PQ\left[\frac{1}{n_1}+\frac{1}{n_2}\right]}}\sim N(0,1)$$

$$Where\ P_1=\frac{x_1}{n_1}\ ,\ P_2=\frac{x_2}{n_2},\ P=\text{probability of Success}=\frac{x_1+x_2}{n_1+n_2},\ Q=1-P$$

Step 4: The Z-table value for a specified level significance taken from normal tables.

Step 5: Conclusion: Since $Z_{cal.val} \leq Z_{table.\ value}$, we accept our H_0 at α % L.O.S and hence we conclude that the two proportions are equal. If $Z_{cal.val} > Z_{table.\ value}$, We reject our H_0 at α % L.O.S and hence we conclude that the two proportions are not equal.

12.7 INFERENCE GREATER THAN OR EQUAL TO TWO PROPORTIONS:

Suppose we want to test whether two or more than two binomial populations have the same parameter p. Referring to these parameters as P_1, $P_{2,...}P_k$. The null hypothesis will be $H_0: P_1=P_2=P_3=\ldots=P_k=P$ and the alternative hypotheses is that these population proportions are not equal.

To test this, we require independent random samples of sizes $n_1,n_2,\ldots n_k$ from k-populations. Suppose there are k successes $x_1,x_2,\ldots x_k$ out of K-populations with sample of sizes $n_1,n_2,\ldots n_k$ is given by the test statistic

$$Z_i = \frac{x_i - n_i p_i}{n_i p_i(1-p_i)} \sim N(0,1)$$

Which are approximately the standard normal distribution. But we know that the square of a random variable having the chi-square distribution with one degree of freedom. The sum of k independent random variables having Chi-square with k degrees of freedom. Then

$$\chi^2 = \sum_{i=1}^{k} Z_i^2 = \sum_{i=1}^{k} \frac{\left(x_i - n_i p_i\right)^2}{n_i p_i(1-p_i)}$$

is a χ^2 -distribution with v degrees of freedom. As $P_1=P_2=P_3=\ldots.=P_k=P$ from null hypothesis are all equal, therefore

$$estimate \quad p = \frac{x_1 + x_2 + \ldots. + x_k}{n_1 + n_2 + \ldots. + n_k}$$

null hypothesis should be rejected if the difference x_i and $n_i p_i$ are large. And the critical region is $\chi^2 > \chi^2_\alpha$ where χ^2_α is and the number of degrees of freedom is k-1. The loss of one degree of freedom results from substituting P, the estimate.

We compare two or more sample proportions which is convenient to determine the value of the χ^2 statistic by looking at the data as arranged in the following way.

	Sample 1	Sample 2		Sample k	Total
Successes	x_1	x_2		x_k	x
Failures	n_1-x_1	n_2-x_2		n_k-x_k	n-x
Total	n_1	n_2		n_k	n

The entry in the cell belonging to the i^{th} row and j^{th} column is called the observed cell frequency O_{ij} i = 1, 2 ; j = 1, 2 ,...,k

under the null hypothesis $P_1 = P_2=...=P_k =P$

$$P = \frac{Total\ number\ of\ Successes}{Total\ number\ of\ trails}$$

$\therefore$ Expected number of successes and failures for the j^{th} Sample one estimated by

$$e_{1j} = n_j p = n_j \frac{x}{n} \quad and \quad e_{2j} = n_j (1-p) = \frac{n_j (n-x)}{n}$$

The quantities e_{1j} and e_{2j} are called the expected call frequencies for j = 1, 2, - - k

Note: The expected frequency for any given call may be obtained by multiplying the totals of the column and the row to which it belongs and then dividing the ground total n .

The test-statistic i.e concerning difference among proportion

$$\chi^2 = \sum_{i=1}^{2} \sum_{j=1}^{k} \frac{\left(O_{ij} - e_{ij}\right)^2}{e_{ij}}$$

12.8 WORKED OUT EXAMPLES:

Example 1: A study shows that 16 of 200 tractors produced on one assembly line required extensive adjustments before they could be shipped. While the same was true for 14 of 400 tractors produced on another assembly line. At the $\alpha = 1\%$ level of significance, does this support the claim that the second production line does superior work.

Solution:

Step 1: The Null hypothesis is $H_0: P_1 = P_2$ i.e, there is no significant difference between two proportions.

Step 2: The alternative hypothesis is $H_1: P_1 \neq P_2$ i.e, there is no significant difference between two proportions. Then $H_1: P_1 < P_2$ right tailed test.

Step3: The test statistic to test the null hypothesis H_0 against the alternative hypothesis H_1 is given by Z-test statistic for 2 proportions for large sample.

$$Z = \frac{(P_1 - P_2)}{\sqrt{PQ\left[\frac{1}{n_1} + \frac{1}{n_2}\right]}} \sim N(0,1)$$

$$Where\ P_1 = \frac{x_1}{n_1}$$

$x_1 = No.of\ tractors\ on\ first\ assembly\ line\ which\ require\ adjustments = 16$

$n_1 = Sample\ size\ of\ \ first\ sample = 200$

$then\ P_1 = \frac{x_1}{n_1} = \frac{16}{100} = 0.08$

$P_2 = \frac{x_2}{n_2}$

$x_2 = No.of\ \ tractors\ on$ second $assembly\ line\ which\ require\ adjustments = 14$

$n_2 = Second\ \ Sample\ size = 400$

$Then\ P_2 = \frac{14}{100} = 0.035$

P *is given by* $P = \frac{x_1 + x_2}{n_1 + n_2}$, (The probability of Success)

$Q\ the\ probability\ of\ \ failure = 1 - P = 1 - \frac{x_1 + x_2}{n_1 + n_2} = \frac{(n_1 + n_2)(x_1 + x_2)}{n_1 + n_2}$

$\therefore P = \frac{x_1 + x_2}{n_1 + n_2} = \frac{16 + 14}{200 + 400} = 0.05$

$Q = 1 - P = 1 - 0.05 = 0.95$

$$Then\quad Z_{cal.value} = \frac{(P_1 - P_2)}{\sqrt{PQ\left[\frac{1}{n_1} + \frac{1}{n_2}\right]}} \sim N(0,1) = \frac{0.080 - 0.035}{\sqrt{0.05 \times 0.95\left[\frac{1}{200} + \frac{1}{400}\right]}} = 2.37$$

Step 4: The Z table or critical value is Z_α = 2.33 for a right-tailed test at 5% L.O.S. taken from normal tables.

Step 5: Conclusion

Since $Z_{cal.value}$ = 2.37 > $Z_{table\ value}$ =2.33 at α = 5% Level of significance we reject our null hypothesis H_0 and hence we conclude that there is significant different between the two proportion.

Example 2:

The owner of a machine shop must decide which of two Snack vending machines to install in his shop. If each machine is Tested 250 times. The first machine fails to work 13-times and the second machine fails to work 7-times. Test at α = 5% level of significance whether the difference between the Corresponding sample proportions is significant.

Solution: Step1: The null hypothesis is Ho: $P_1=P_2$, i.e, there is no Significant difference between the sample proportions.

Step 2: The alternative hypothesis is $H_1: P_1 \neq P_2$ i.e, there is no significant difference between the two sample proportions is a two tailed test.

Step3: The test statistic to test the above hypothesis H_0 Vs H_1 is given by

Test statistic $$Z = \frac{(P_1 - P_2)}{\sqrt{PQ\left[\frac{1}{n_1} + \frac{1}{n_2}\right]}} \sim N(0,1)$$

x_1 *is the no.of times first mcachine fails* $= 13$

n_1 *is the size of the first sample* $= 250$

x_2 *is the No.of times* second *machine fails* $= 7$

$$\therefore P_1 = \frac{x_1}{n_1} = \frac{13}{250} = 0.052 \, , P_2 = \frac{x_2}{n_2} = \frac{7}{250} = 0.028$$

$$\therefore P = \frac{x_1 + x_2}{n_1 + n_2} = P = \frac{13+7}{250+250} = \frac{20}{500} = 0.04$$

$$Q = 1 - P = 1 - 0.04 = 0.96$$

$$Z_{cal.value} = \frac{0.052 - 0.028}{\sqrt{0.04 \times 0.96\left[\frac{1}{250} + \frac{1}{250}\right]}} \sim N(0,1) = 1.37$$

Step 4: The Z table or critical value is $Z_{\alpha/2} = 1.96$ at $\alpha = 5\%$ L.O.S. taken from normal tables.

Step 5:<u>Conclusion</u>

Since $Z_{cal.value} = 1.37 < Z_{table\ value} = 1.96$ we accept our null hypothesis at $\alpha = 5\%$ Level of significance(L.O.S). Hence we conclude that there is no significant difference between the sample proportions.

Example 3:

Photo lithography plays a central role in manufacturing integrated circuits made on thin discs of silicon prior to a quality improvement program. Too many rework operations were required. In a sample of 200 units, 26 required reworking of the photo lithographic step. following training in the use of pareto chats and other approaches to identify significant problems, improvements were

made. A new sample of size 200 had only 12 that needed rework. Find the a Large Sample 99% confidence interval for the difference of the true proportions.

Solution: Large sample confidence interval for the difference of two proportions is

$$(P_1 - P_2) \pm Z_{\alpha/2}\sqrt{PQ\left[\frac{1}{n_1}+\frac{1}{n_2}\right]} \text{ and } (P_1 - P_2) \pm Z_{\alpha/2}\sqrt{\frac{P_1Q_1}{n_1}+\frac{P_2Q_2}{n_2}}$$

$$P_1 = \frac{x_1}{n_1} = \frac{26}{200} = 0.13, P_2 = \frac{x_2}{n_2} = \frac{12}{200} = 0.6$$

$Z_{\alpha/2} = 2.58$ *is the Z.critical value taken from normal tables.*

$$Q_1 = 1 - P_1 = 1 - 0.13 = 0.87$$

$$Q_2 = 1 - P_2 = 1 - 0.03 = 0.94$$

Then we have

$$P_1 - P_2 \pm Z_{\alpha/2} = \sqrt{\frac{P_1(1-P_1)}{n_1}+\frac{P_2(1-P_2)}{n_2}}$$

$$= 0.13 - 0.06 \pm 2.58\sqrt{\frac{0.13\times 0.87}{200}+\frac{0.06\times 0.94}{200}}$$

$$= 0.13 - 0.06 \pm 2.58 \times 0.029 = 0.07 \pm 0.075$$

That is; =(0.07-0.075, 0.07+0.075)

Confidence interval=(-0.005, 0.145)

Example 4: Some defendants in criminal proceedings plead guilty and are sentenced without a trial, whereas others who plead Innocent are subsequently found guilty and then are sentenced. In recent years, legal scholars have speculated as to whether sentences of those who plead guilty differ in severity from sentences for those who plead innocent and one subsequently judged guilty. Consider the accompanying data on defendants from San Francisco county accused of robbery, all of whom had previous prison rewards. Does this data suggest that the proportion of all defendants in these circumstances who plead guilty and are sent to prison differs from the proportion who are sent to prison after pleading innocent and being found guilty?

Plea		
	Guilty	**Not guilty**
Number judged guilty	**n1=191**	**n2=64**
Number judged guilty	**x1=101**	**x2=56**
Sample proportion	p_1	p_2

Solution: Step1: Let P_1 and P_2 denote the two population proportions, The null hypothesis of interest are Ho: P_1-P_2=0, i.e, there is no Significant difference between two proportions.

Step 2: The alternative hypothesis is H_1:$P_1 - P_2 \neq 0$ i.e, there is no significant difference between two proportions.

Step3: The test statistics to test the above hypothesis H_0 Vs H_1 is given by Z-test for large n.

i.e

Test statistic $$Z = \frac{(P_1 - P_2)}{\sqrt{PQ\left[\frac{1}{n_1} + \frac{1}{n_2}\right]}} \sim N(0,1)$$

Where

$$P_1 = The\ sample\ proportion\ from\ 1st\ sample = 0.529 = \frac{x_1}{n_1}$$

$$P_2 = The\ sample\ proportion\ from\ 2nd\ sample = 0.875 = \frac{x_2}{n_2}$$

$$P - The\ combined\ estimate\ of\ the\ common\ success\ proportion = \frac{x_1 + x_2}{n_1 + n_2} = \frac{101 + 56}{191 + 64} = 0.616$$

$$Q = 1 - P_1 = 1 - 0.616 = 0.384$$

$$Then\quad Z_{cal.value} = \frac{0.529 - 0.875}{\sqrt{0.616 \times 0.384\left[\frac{1}{191} + \frac{1}{64}\right]}} \sim N(0,1)$$

$$= \frac{-0.346}{0.070}$$

$$Z = -4.94$$

$$|Z| = 4.94$$

Step 4: The $Z_{\text{table value}}$ at 5% L.O.S. is $Z_{0.05}$=2.58 from normal tables.

Step 5: Since $Z_{cal.value} = 4.94 > Z_{table\ value} = 2.58$ at $\alpha = 5\%$ Level of significance(L.O.S). We reject our null hypothesis and Hence we conclude that there is significant difference between the two proportions.

The P-value for a two-tailed Z-test is

P-value=2[1- Φ |Z|]=2[1- Φ (4.94)]<2[1- Φ (3.49)]=0,0004

An extensive standard normal table yields P-value = 0.0000006 this P-value is so minute that at any reasonable level α H_0 should be rejected. the data very strongly suggest that $P_1 \neq P_2$ and in particular, that initially pleading guilty may be a good strategy as far as avoiding prison is concerned

Example 5: If for one half of n events, the chance of success is p and chance of failure is q. whilst for the other half, the chance of success is q and the chance of failure is p. show that the standard deviation of the number of successes is the same as if the chance of successes were p in all the cases i.e $\sqrt{npq}$

But the mean of number of successes is n/2 and not np.

Solution: Let the number of successes in first and second half be denoted by x and y respectively.

$$E(x) = \frac{np}{2}, E(x) = \frac{nq}{2}$$

$$V(x) = \frac{1}{2}\text{npq}, v\text{ar(y)} = \frac{1}{2}\text{npq}$$

$$E(\text{x+ y}) = \text{E(x)} + \text{E(y)} = \frac{np}{2} + \frac{nq}{2} = \frac{n}{2}(p+q) = \frac{1}{2}n$$

$$V(x+y) = Var(x) + Var(y) = \frac{1}{2}npq + \frac{1}{2}npq = npq$$

Example 6:

In two large populations there are 30 and 25 percent respectively fair haired people. Is this difference likely to be hidden in samples of 1200 and 900 respectively from the two populations?

Solution:

$$Here\ p_1 = \frac{30}{100} = 0.30\ ,\quad P_2 = \frac{25}{100} = 0.25$$

$$so\ that\ p_1 - P_2 = 0.05$$

$$E^2 = \frac{p_1 q_1}{n_1} + \frac{p_1 q_1}{n_1} = \frac{0.30 \times 0.70}{100} + \frac{0.25 \times 0.75}{900} = 0.000175 + 0.00021 = 0.000383$$

$$E = \sqrt{0.000383} = 0.0195$$

$$\therefore Z = \frac{p_1 - P_2}{E} = \frac{0.05}{0.0195} = 2.56$$

Hence it is unlikely that real difference will be hidden

Example 7:

In a random sample of 500 men from a particular district of U.P. 300 are found to be smokers. In one of 1000 men from another district, 550 are smokers. Do the data indicate that the two districts are significantly different with respect to the prevalence of smoking among men?

Solution:

$$Here \;\; p_1 = \frac{300}{500} = 3/5, \quad P_2 = \frac{550}{1000} = 11/20$$

$$p_0 = \frac{p_1 n_1 + p_2 n_2}{n_1 + n_2} = \frac{300 + 550}{500 + 1000} = \frac{17}{30}, \;\; q_0 = 1 - p_0 = 1 - \frac{17}{30} = \frac{30 - 17}{30} = \frac{13}{30}$$

$$E = \sqrt{p_0 q_0 \left[\frac{1}{n_1} + \frac{1}{n_2}\right]} = \sqrt{\frac{17}{30} \times \frac{13}{30}\left[\frac{1}{500} + \frac{1}{1000}\right]} \;\; = 0.0271$$

$$p_1 - p_2 = 0.6 - 0.55 = 0.05$$

$$\therefore Z = \frac{p_1 - p_2}{E} = \frac{0.05}{0.0271} = 1.9 \; (Approximately)$$

Hence the difference is not significant i.e, the data do indicate that the two districts are significantly different with respect to the prevalence of smoking men.

Example 8:

If for one half of n events the chance of success is p and the chance of failure is q, whilst for the other half the Chance of success is q if and the chance of failure is p. Show that the standard deviation of the member of success is the same as if the chance of success were p in all the cases i.e $\sqrt{npq}$ but that the mean of the number of success is n/2 and not np.

Solution: Let σ_1 *and* σ_2 of and of denote the standard deviation of first and second halves of n events.

Thus, $\sigma_1^2 = \frac{1}{2}\text{npq}$ *and* $\sigma_2^2 = \frac{1}{2}\text{npq}$

Hence, the standard deviation of the number of successes.

$$\sqrt{\sigma_1^2 + \sigma_2^2} = \sqrt{\left[\frac{n}{2}\text{pq} + \frac{n}{2}\text{qp}\right]} = \sqrt{npq}$$

If P_0 denotes the proportion of success in the n-events,

$$\text{Then } P_0 = 1 - \frac{\frac{n}{2}p + \frac{n}{2}q}{\frac{n}{2} + \frac{n}{2}} = \frac{1}{2}(p + q) = \frac{1}{2}$$

Hence, the mean of the number of success=$nP_0 = \frac{1}{2}n$

Example 9:

Tests are made on the proportion of defective castings produced by five different moulds. If there were 14 defectives among 100 castings made with mould-I. 33 defectives among 200 castings made with mould-II. 21 defectives among 180 castings made with mould-III. 17 defectives among 120 castings made with mould-IV and 25 defectives among 150 castings made with mould-V. Use the

α=1%. level of significance to test whether The true proportion of defectives is the same for each mould.

Solution:

Step 1:The null hypothesis is H_0:$P_1=P_2=P_3=P_4=P_5$ i.e all the proportions are equal.

Step 2: The alternative hypothesis is H_1:$P_1\neq P_2\neq P_3\neq P_4\neq P_5$ all are not equal.

Step 3:The information given in problem are tabulated

	Mould-I	Mould-II	Mould-III	Mould-IV	Mould-V	Total
Defectives	14	33	21	17	25	10
Non Defectives	86	167	159	103	125	640
Total	100	200	180	120	150	750

The test statistic to test the above hypothesis is given by

$$\sum_{i=1}^{2}\sum_{j=1}^{5}\frac{\left(O_{ij}-e_{ij}\right)^2}{e_{ij}}\sim\chi^2_{k-1}\ d.t$$

The expected frequencies

$$e_{11}=\frac{110\times100}{750}=14.67$$

$$e_{12}=\frac{110\times200}{750}=29.37,\ e_{13}=\frac{110\times180}{750}=26.4$$

$$e_{14}=\frac{110\times120}{750}=17.6,\ e_{15}=\frac{110\times150}{750}=21.96$$

$$e_{21}=\frac{640\times100}{750}=85.33,\ e_{22}=\frac{640\times200}{750}=170.63$$

$$e_{23}=\frac{640\times180}{750}=153.6,\ e_{24}=\frac{640\times120}{750}=102.4$$

$$e_{25}=\frac{640\times150}{750}=128.04$$

$$The\chi^2-test\ statistic=\sum_{i=1}^{2}\sum_{j=1}^{5}\frac{\left(O_{ij}-e_{ij}\right)^2}{e_{ij}}\chi^2_{k-1}\ d.t$$

$$Hence\ \chi^2_{cal.value} = \frac{(14-14.67)^2}{14.67} + \frac{(33-29.37)^2}{29.37} + \frac{(21-26.4)^2}{26.4} + \frac{(17-17.6)^2}{17.6} + \frac{(25-21.96)^2}{21.96}$$

$$+ \frac{(86-85.33)^2}{85.33} + \frac{(167-170.63)^2}{170.63} + \frac{(159-153.6)^2}{153.6} + \frac{(103-102.4)^2}{102.4} + \frac{(125-128.04)^2}{128.04} = 2.37$$

Step 4: For (k-1)d.t=(5-1)=4 d.t

$Hence\ \chi^2_{-\text{table}\ value}\ \ at\ \alpha = 1\%\ \ L.O.S\ \ is\ \ \ \chi^2_{0.01} = 13.277$

Step 5: Since $\chi^2_{cal.\,value} = 2.37 < \chi^2_{\text{table}\ value} = 13.277$

We accept our null hypothesis at α= 1%. L.O.S for (k-1)=4 d.t. Hence we conclude that all the proportions are equal.

12.9 EXERCISE:

1) In a sample of 600 students of a certain college 400 are found to use ball pens. In another college from a sample of 900 students 450 were found to use ball pens. Test whether 2 colleges are Significantly different with respect to the habit of using ball pens.

2) During a country wide investigation the incidence of tuberculosis was found to be l%. In a college of 400 strength 5 reported to be affected whereas in another 1200 strength 10 were affected a) Does this indicate any significant difference. b) If the population proportion of tuberculosis is not known test whether the difference is significant.

3) A candidate for election made a speech in city A but not in B. A sample 500 voters from city A showed that 59.6% of the votes were in favour of him where as a sample of 300 voters from city B showed that 50% of the voters favoured him. Discuss whether his speech could product any effect on voters in city A use 5% L.O.S.

4) Random Samples of 400 men and 600 women in a locality were asked whether they would like to have a bus stop near their residence. 200 men and 325 women were in favour of the proposal. Test the hypothesis that proportions of men and women in favour of proposals are same in the male and female. Test at 5% L.O.S.

5) The following data come from a study in which random samples of the employees of three government agencies were asked questions about their pension plan.

	Agency 1	Agency 2	Agency 3
For the pension plan	67	84	109
Against the pension plan	33	66	41

Use the $\alpha = 1\%$ L.O.S to test the null hypothesis that the actual proportions of employees favoring the pension plan are the same.

12.10 SUMMARY

In this lesson an attempt is made to explain the concepts of inference on two or more proportions, estimation of proportions, approximations and procedures associated with them doing with both theory and practical. A number of examples are worked out and a good number of exercises are also given.

12.11 TECHNICAL TERMS

- Inference concerning two proportions
- Inference concerning more than two proportion
- Z-test statistic
- χ^2-Test statistic

12.12 SELF ASSESSMENT QUESTIONS

SHORT:

1. What is the formula for the confidence interval of the difference between two proportions?
2. How is the margin of error (E) calculated for two proportions?
3. What is the pooled proportion in hypothesis testing for two proportions?
4. When is the Chi-Square test used for comparing multiple proportions?
5. What are the null and alternative hypotheses for testing the equality of two proportions?

ESSAY:

1. Explain the concept of confidence intervals for the difference between two proportions and their significance in statistical inference.
2. Discuss the role of sample size in estimating two proportions and how it affects the margin of error.
3. Describe the hypothesis testing procedure for comparing two proportions, including the use of the pooled proportion.
4. Compare and contrast hypothesis testing for two proportions versus three or more proportions.
5. Explain the Chi-Square test for homogeneity and its applications in analyzing multiple proportions.

12.13 FURTHER READINGS

1 "A First Course in Probability" – Sheldon Ross
2 "Introduction to Probability" – Dimitri P. Bertsekas & John N. Tsitsiklis
3 "Probability and Statistics" – Morris H. DeGroot & Mark J. Schervish
4 "Probability and Random Processes" – Geoffrey Grimmett & David Stirzaker
5 "Probability: Theory and Examples" – Rick Durrett

Dr. T.V. Pradeep Kumar

Lesson-13

COMPARING TWO MEANS

OBJECTIVES:

1. To understand the logical framework for estimating the difference between the means of two distinct populations and performing tests of hypothesis Concerning those means.
2. To learn how to perform a test of hypothesis Concerning the difference between the means of two distinct populations using large, independent samples.

STRUCTURE

13.1 Introduction

13.2 Independent and dependent Samples

13.3 Testing of hypothesis

13.4 t-test for single mean

13.5 t-test for Difference of means of two small Samples

13.6 Paired T-Test For Difference Of Means:

13.7 Solved problems

13.8 Practice Problems

13.9 Exercise

13.10 Summary

13.11 Technical Terms.

13.12 Self Assessment Questions

13.13 Further Readings

13.1 INTRODUCTION

Comparing two means is a statistical process that uses sample data to determine if the means of two groups are different. It is often used in testing hypothesis. Here we may select if the test will be one-sample, two-sample or paired samples. Then decide if the test will be one-tailed or two tailed. Then Set the null hypothesis and the alternative hypothesis. Calculate the z-statistic or t-statistic. First we need to consider whether the two populations are independent. When considering the sample mean, there were two parameters we had to consider, the population mean and S.D. In second Step, we determine where the population standard deviations are the same or different.

13.2 INDEPENDENT AND DEPENDENT SAMPLES

13.2.1 Definition: The Samples from two populations are said to be independent if the samples selected from one of the populations has no relationship with the samples selected from the other population.

13.2.2 Definition: The samples from thud populations are said to be dependent (or) paired data each measurement in one sample is matched or paired with a particular measurement in the other sample.

13.2.3 Remark: Another way to Consider the above is how many measurements are taken off of each subject.

If we take only one measurement, then it is independent.

If we take two measurements, then they are paired data.

Note that the exceptions are in familiar situations such as in a study of spouses or twins. In such cases the data is almost always treated as dependent (or) paired data Samples.

13.2.4 Example: We want to compare the gas mileage of two brands of gasoline. Describe how to design a study Involving.

Sol: (a) Independent sample:

Randomly assign 12 cars to use brand A and another 12 cars to use brand B.

(b) Dependent samples:

Using 12 cars, have each car use brand A and Brand B.

Then Compare the differences in mileage for each car.

(Try)

13.2.5. Note: The two types of samples require a different theory to construct a confidence interval and then develop a hypothesis test.

13.3 TESTING OF HYPOTHESIS

Hypothesis concerning the relative sizes of the means of two populations are tested closing the same critical value and p-value procedures that were used in the case of a single population.

All that is needed is to know how to express the null and alternative hypotheses and to know the formula for the standard test statistic and the distribution that it follows:

The null and Alternative hypothesis will always be expressed interns of the difference of the two population means.

So, the null hypothesis will always be written as $H_0: \mu_1 - \mu_2 = D_0$ (say)
Where D_0 is a number that is deduced from the statement of the situation.

As in the case of a single population the alternative hypothesis can take one of the following three forms:
(a) $H_1: \mu_1 - \mu_2 < \mu_0$ (left tailed)
(b) $H_1: \mu_1 - \mu_2 > \mu_0$ (Right tailed)
(c) $H_1: \mu_1 - \mu_2 \neq \mu_0$ (Two-tailed)

13.3.1 Standardized test: The tests statistic for hypothesis concerning the difference between two population means (for large independent samples)

The test formula is

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - \mu_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Where $\bar{x}_1, \bar{x}_2$ are means, of populations 1 and 2 , n_1, n_2 are sizes of populations.
13.3.2 Note: The test statistic has the Standard normal distribution. The samples must be independent, and each sample must be large,

$$\text{i.e. } n_1 \geq 30, n_2 \geq 30$$

13.3 Example: The means of two single large Samples of 1000 and 2000 members are 67.5 inches and 68.0 inches respectively. Can the samples be regarded as drawn from the same population of standard deviation 2.5 inches? (Test at 5% level of Significance)

Solution: In usual notations, we write the given data by the above hypothesis is as follows:

$$n_1 = 1000, n_1 = 2000, \overline{x_1} = 67.5 \text{ inches}, \overline{x_2} = 68.0 \text{ inches}$$

Now, write hypothesis as:
Null hypothesis: $H_0 : \mu_1 = \mu_2$ *and* $\sigma = 2.5$ inches that is, the samples have been drawn from the same population and S.D. 2.5 inches.
Alternative hypothesis: $H_1 : \mu_1 \neq \mu_2$ (Two tailed).

Then, the test the statistic, under H_0.The test statistic is:

$$Z = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0,1)$$ (Since Size of the sample Large)

$$= \frac{67.5 - 68.0}{2.5\sqrt{\left(\frac{1}{1000} + \frac{1}{2000}\right)}} = \frac{-0.5}{2.5 \times 0.0387} = -5.1$$

Conclusion:

Since |Z|>3, the value is highly significant and we reject the null hypothesis and conclude that Samples are not from the same population and Standard deviation 2.5

Note: How ever its sample sizes are small, then an exact Sample test, t-test for difference of means is to be used.

13 .4. T-TEST FOR SINGLE MEAN

Now, to test whether the mean of a sample drawn from a normal population deviates Significantly from a Standard Value when Variance of the population is unknown.

Here, Null hypothesis

H_0 :There is no Significant difference between the sample mean $\bar{X}$ and the population mean′μ' that is, we use the statistic

$$t = \frac{\bar{X} - \mu}{s/\sqrt{n}}, \text{Where } \bar{X} \text{ is the mean of the sample.}$$

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

with (n-1) degrees of freedom.

Note that, at given level of significance of α_1 and (n-1) degrees of freedom.

Here refer t-table for t_α (one tailed or two tailed).

If calculated t-value is such that $|t| < t_\alpha$,

the null hypothesis (H_0) is rejected.

13.4.1 Limits of population mean:

(1) If t_α is the table of 't' at level of Significance 'α' at (n - 1) degrees of freedom

(2) for acceptance of H_0,

$$\left|\frac{\bar{X}-\mu}{s/\sqrt{n}}\right| < t_\alpha \qquad \rightarrow (A)$$

$$Where \quad \bar{x} - t_\alpha \, s/\sqrt{n} \; < \; \bar{x} + t_\alpha \, s/\sqrt{n}$$

(3) 95%. Confidence limits (Level of significance 5%) are $\bar{X} \pm t_{0.005} \, (s/\sqrt{n})$

(4) 99%. Confidence limits

13.4.2.Example: A Sample of 20 items has mean 42 units and S.D. 5 units. Test the hypothesis that it is a random sample from a normal population with mean 45 Units.

Solution: from the given hypothesis,

write H_0: There is no significant difference between sample mean and the population mean.

That is $\mu = 45 \, units$

H_1: $\mu \neq 45$(Two tailed test)

$$Given \; that \; n = 20, \bar{X} = 42, s = 5, \; \gamma = 19 \; d.f.$$

$$s^2 = \frac{n}{n-1} s^2 = \left[\frac{20}{20-1}\right](5)^2 = 26.31,$$

$$\therefore s = 5.129$$

$$By \; t-test, \; t = \left|\frac{\bar{X}-\mu}{s/\sqrt{n}}\right| = \frac{42-45}{5\left(\frac{120}{\sqrt{20}}\right)}$$

$$= -2.615, \quad |t| = 2.615$$

The tabulated value of 't' at 5% level for 19 df is $t_{0.05}$=2.09

Conclusion: Since $|t|>t_{0.05}$, The hypothesis H_0 is rejected, that is there is significant difference between the Sample mean and population mean.

13.5 T-TEST FOR DIFFERENCE OF MEANS OF TWO SMALL SAMPLES (FROM A NORMAL POPULATION)

This test is ward to test whether the two samples $x_1, x_2, \ldots x_{n_1}$, $y_1, y_2, \ldots y_{n_2}$ of sizes n_1 and n_2 have been drawn from two normal populations with means μ_1 *and* μ_2 respectively under the assumption that the population variance are equal $(i.e\ \sigma_1 = \sigma_2 = \sigma)$

Here

H_0: The samples have been drawn from the normal population with means μ_1 *and* μ_2

i.e $H_0 : \mu_1 \neq \mu_2$

Let $\bar{X}, \bar{Y}$ be their means of the two Samples under this null hypothesis H_0, the test of static 't' is given by

$$t = \frac{\bar{X} - \bar{Y}}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} < t\left(n_1 + n_2 - 2\ d.f\right) \qquad \rightarrow (A)$$

13.5.1 Remarks:

(1) If the two sample S.D's s_1, s_2 are given then we have

$$s^2 = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}$$

If $n_1 = n_2 = n$, $t = \dfrac{\bar{X} - \bar{Y}}{\sqrt{\dfrac{s_1^2 + s_2^2}{n-1}}}$ *can be used as a test statistic.*

13.5.2. Example: **Two samples of Sodium bulbs were tested for length of life and the following results were got:**

	Size	Sample mean	Sample S.D
Type-1	**8**	**1234 hrs**	**36 hrs**
Type-2	**7**	**1036 hrs**	**40 hrs**

Is the difference in the means significant to generalize that type-1 is superior to type 2 regarding length of life.
Solution:

Here

$H_0 : \mu_1 = \mu_2$, i.e two types of bulbs have Same life time.

$H_0 : \mu_1 > \mu_2$ i.e two type 1 is superior to type 2.

$$\Rightarrow s^2 = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}$$

$$= \frac{8 \times (36)^2 + 7(40)^2}{8 + 7 - 2}$$

=1659.076

$\therefore s = 40.7317$

Now the t-static,

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{1234 - 1036}{40.7317\sqrt{\frac{1}{8} + \frac{1}{7}}}$$

$= 18.1480 \sim t(n_1 + n_2 - 2\, d.\text{f})$

$t_{0.05}$ at d.f 13 is 1.77 (one tailed test)

Conclusion: Since calculated $|t| > t_{0.05}$ H_0 is rejected i.e H_1 is accepted.

∴ Type 1 definetly superior to type 2

Where $\bar{X} = \sum_{i=1}^{n_1} \frac{X_i}{n_i}$, $\bar{Y} = \sum_{j=1}^{n_2} \frac{Y_j}{n_j}$

$$s^2 = \frac{1}{n_1 + n_2 - 2}\left[\sum (X_i - \bar{X})^2 + (Y_j - \bar{Y})^2\right]$$

Is an unbiased Variance σ^2

$\Rightarrow t$ *follows* $t-$*distribution* $n_1 + n_2 - 2$ *d.f*

13.6 PAIRED T-TEST FOR DIFFERENCE OF MEANS:

Now we Consider the Case

(i) The Sample sizes are equal, i.e $n_1 = n_2 = n(say)$

(ii) The two samples are not Independent but the sample observations are paired together

i.e the pair of observations

(x_i, y_i) , $i = 1, 2, \ldots n$ corresponds to the same(i^{th}) *sample unit*.

The problem is to test the sample means to differ significantly or not.

The paired t-test is as follows

Here we consider the increments $d_i = x_i - y_i$, $i = 1, 2, \ldots n$

Under the null hypothesis, H_0 That increments are due to fluctuations of Sampling,

$$t = \frac{\bar{d}}{s/\sqrt{n}}$$

where $\bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i$ *and* $s^2 = \frac{1}{n-1}\sum_{i=1}^{n} (d_i - \bar{d})^2$

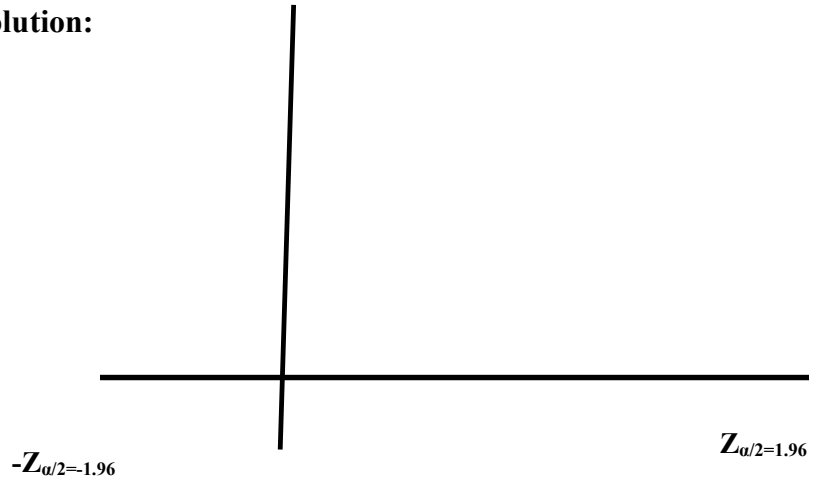
which follows Student t-distribution with (n-1) d f.

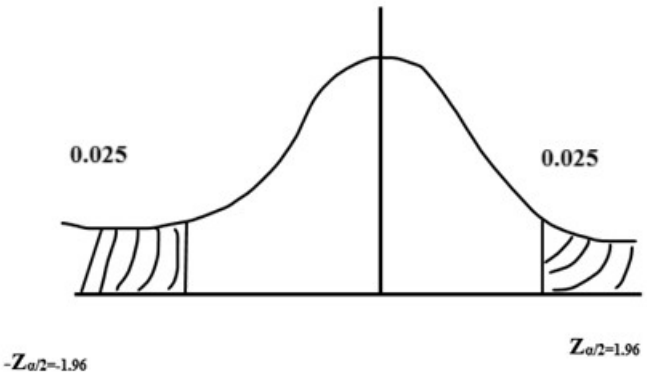
13.7 SOLVED PROBLEMS

1) Samples of students were drawn from two Universities and from their weights in kgs and S.D are calculated, make a large sample test to test the Significance of the the difference between the means.

	Mean	S.D	Size of the sample
University A	**55**	**10**	**400**
University B	**57**	**15**	**100**

Solution:





Here (a) Null hypothesis $H_0 : \mu_1 = \mu_2$

(b) Alternative hypothesis $H_1 : \mu_1 \neq \mu_2$

(c) Level of Significance α=0.05

(d) Critical region:

$$Z_{\frac{\alpha}{2}} = 1.96 \; for \; \alpha = 0.05 \; level \; of \; significance$$

e) Test of Statistic.

$$Z = \frac{(\overline{x_1} - \overline{x_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Here $\overline{x_1} = 55$, $\overline{x_2} = 57$

$s_1 = 10 \Rightarrow s_1^2 = 10^2 = 100$

$s_2 = 15 \Rightarrow s_2^2 = 15^2 = 225$

$n_1 = 400, n_2 = 100$

$$\therefore Z = \frac{(\overline{x_1} - \overline{x_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Note that when two Variances are unknown (σ_1^2, σ_2^2) they can be replaced by Sample Variances

$$Z = \frac{(55-57)-0}{\sqrt{\frac{100}{400} + \frac{225}{100}}} = Z = \frac{-2}{\sqrt{\frac{1}{4} + \frac{9}{4}}} = \frac{-2}{\sqrt{\frac{5}{2}}} = -1.265$$

Conclusion: *Here* $-Z_{\frac{\alpha}{2}} < Z < Z_{\frac{\alpha}{2}}$ *then null hypothesis* H_0 *is accepted*.

- 1.96 < - 1.265 < 1.96 i.e. Null hypothesis is accepted.

There is no Significant difference between the means.

2) A random sample of 16 values from a normal population showed a mean of 41.5 inches and the sum of Squares of deviations from this mean equal to 15 Square inches. Show that the assumption of a mean of 43.5 inches for the population is not reasonable. Obtain 95% and 95% fiducial limits for the same.

[you may use the following Information from statistical tables ϑ=15,

$$\begin{cases} \boldsymbol{P = 0.05}, & t = 2.131 \\ \boldsymbol{P = 0.01}, & t = 2.947 \end{cases}$$

Solution: from the given hypothesis, we have

$n = 16, \overline{x} = 41.5\ inches\ and\ \sum (x - \overline{x})^2\ sq.\ \text{inches}$

$$s^2 = \frac{1}{n-1}\sum (x - \overline{x})^2 = \frac{135}{16-1} = \frac{135}{15} = 9 \Rightarrow s = \sqrt{9} = 3.$$

Now Null hypothesis $H_1 : \mu = 43.5\ inches$. Ho. mu = 43.5 inches

i.e the data is consistent with the assumption that the mean height in the population is 43.5 inches.

Alternative hypothesis: $H_1 : \mu \neq 43.5\ inches$

Test Statistic: H_0, the test statistic is

$$\therefore |t| = \frac{|41.5 - 43.5|}{3/4} = \frac{8}{3} = 2.667$$

Here the number of degrees of freedom is 16-1=15
we are given:$t_{0.05}$ for 15 degrees of freedom = 2.131
and $t_{0.05}$ for 15 degrees of freedom = 2.947

Conclusion: Since Calculated n|t| it is greater than 2.131 null hypothesis rejected at 5%. level of significance and then we conclude that the assumption of mean of 43.5 inches for the population is not reasonable.

3. The means of two random samples of sizes 9.7 are 196.42 and 198.82.the sample variances are 3.375 and 3.17 respectively, Can the samples be drawn from same population?

Solution: $Given\ n_1 = 9, n_2 = 7, \overline{x_1} = 196.42,\ ,\ \overline{x_2} = 198.82, s_1^2 = 3.375, s_2^2 = 3.17$

Since both n_1, n_2 are <30 (small steps) we use t-test.

Then ϑ_1 =n_1-1=8, ϑ_2 =n_2-1=6, are the respective degrees of freedom

$$So\ s^2 = \frac{\vartheta_1 s_1^2 + \vartheta_2 s_2^2}{\vartheta_1 + \vartheta_2} = \frac{8 \times 3.375 + 6 \times 3.17}{8 + 6} = 3.2871$$

$$\Rightarrow s = 1.81$$

Null hypothesis(H_0): $H_0 : \mu_1 = \mu_2$

Alternative hypothesis(H_1): $\mu_1 \neq \mu_2$

Level of significance: α = 0.05

$t_{\alpha/2}\ for\ \vartheta_1 + \vartheta_2\ d\ \text{eg}\ rees\ freedom$

$t_{0.05/2}$ *for* $\vartheta_1 + \vartheta_2$ *degrees freedom*

$t_{0.025}$ *for* $8+6$ *degrees freedom is* 2.145

Test static:

$$t = \frac{(\overline{x_1} - \overline{x_2})}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{196.42 - 198.82}{(1.81)\sqrt{\frac{1}{9} + \frac{1}{7}}} = -2.63$$

|t|=2.63

Conclusion:

$\therefore |t| > t_{\frac{\alpha}{2}}$

$\therefore$ *We reject the null hypothesis.*

13.8 PRACTICE PROBLEM:

Samples of students were drawn from two universities and from their weights in kilograms mean and S.D are calculated and shown below make a large sample examine the significance of difference between means at 1% LOS.

	Mean	**Standard Deviation**	**Sample size**
University A	55	10	10
University B	57	15	20

2. The nicotine in milligrams of two samples of tobacco were found to be as follows. Examine the truth value of the hypothesis for the difference between means at 0.05 level.

Sample-A	24	27	26	23	25	-
Sample-B	29	30	30	31	24	36

Solution: Given $n_1 = 5, n_2 = 6$, Since both n_1, n_2 are < 30 (small samples) we use t-test. then $\vartheta_1 = n_1 - 1 = 4$ and $\vartheta_2 = n_2 - 1 = 5$, are the respective degrees of freedom. Then combined degrees of freedom is $\vartheta = \vartheta_1 + \vartheta_2 = 4 + 5 = 9$.

Null hypothesis ($\mathbf{H_0}$): $H_0: \mu_1 = \mu_2$

Alternative hypothesis ($\mathbf{H_1}$): $\mu_1 \neq \mu_2$

Level of significance: $\alpha = 0.01$
$t_{a/2}$ for ϑ degrees of freedom
$= t_{0.01/2}$ for 9 degrees of freedom
$= t_{0.005}$ for 9 degrees of freedom =3.25.

Calculation table.

X	$\bar{X}$	$X - \bar{X}$	$(X - \bar{X})^2$	Y	$\bar{Y}$	$Y - \bar{Y}$	$(Y - \bar{Y})^2$
24	$\bar{X} = \frac{125}{5} = 25$	-1	1	29	$\bar{Y} = \frac{180}{6} = 30$	-1	1
27		2	4	30		0	0
26		1	1	30		0	0
23		--2	4	31		1	1
25		0	0	24		-6	36
-		-	-	26		6	36
Total = 125			$\Sigma(x - \bar{x})^2 = 10$	180			$\Sigma(y - \bar{y})^2 = 74$

$s_1^2 = \frac{\Sigma(x-\bar{x})^2}{\vartheta_1} = \frac{10}{4} = 2.5$ and $s_2^2 = \frac{\Sigma(y-\bar{y})^2}{\vartheta_2} = \frac{74}{5} = 14.8$

Common variance:

$$S^2 = \frac{\vartheta_1 s_1^2 + \vartheta_2 s_2^2}{\vartheta_1 + \vartheta_2} = \frac{4 \times 2.5 + 5 \times 14.8}{4 + 5} = \frac{84}{9} = 9.5$$
$$\Rightarrow S = \sqrt{9.5} = 3.09$$

Test statistic:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{25 - 30}{(3.09)\sqrt{\frac{1}{5} + \frac{1}{6}}} = \frac{-5}{1.88} = -2.65$$

$$|t| = 2.65$$

Conclusion: since $|t| < t_{a/2}$, We accept the Null hypothesis

3. Two types of drugs were used on 5 and 7 patients for reducing their weight. Drug A was imported and drug B indigenous. The decrease in the weight after using the drugs for six months were as follows:

Drug A:	10	12	13	11	14	-	-
Drug B:	8	9	12	14	15	10	9

At 5% LOS can we believe that drug A is more effective than drug B ?

Solution: Given $n_1 = 5, n_2 = 7$, Since both n_1, n_2 are < 30 (small samples) we use t-test. then $\vartheta_1 = n_1 - 1 = 4$ and $\vartheta_2 = n_2 - 1 = 6$, are the respective degrees of freedom. Then combined degrees of freedom is $\vartheta = \vartheta_1 + \vartheta_2 = 4 + 6 = 10$.

Null hypothesis (H_0): $H_0: \mu_1 = \mu_2$

Alternative hypothesis (H_1): $\mu_1 > \mu_2$

It is a one tailed test so we have to use t_d as table value.

Level of significance: $\alpha = 0.05$
t_d for ϑ degrees of freedom
$= t_{0.05}$ for 10 degrees of freedom
$= t_{0.05}$ for 10 degrees of freedom $= 1.182$

Calculation table.

X	$\bar{X}$	$X - \bar{X}$	$(X - \bar{X})^2$	Y	$\bar{Y}$	$Y - \bar{Y}$	$(Y - \bar{Y})^2$
10	$\bar{X} = \frac{60}{5} = 12$	2	4	8	$\bar{Y} = \frac{77}{7} = 11$	-3	9
12		0	0	9		-2	4
13		+1	1	12		+1	1

11		-1	1	14		+3	9
14		+2	4	15		+4	16
		-	-	10		-1	1
				9		-2	4
Total $= 60$			$\Sigma(x-\bar{x})^2 = 10$	77			$\Sigma(y-\bar{y})^2 = 44$

$$s_1^2 = \frac{\sum (x-\bar{x})^2}{\vartheta_1} = \frac{10}{4} = 2.5 \text{ and } s_2^2 = \frac{\sum (y-\bar{y})^2}{\vartheta_2} = \frac{44}{6} = 6.33$$

Common variance:

$$S^2 = \frac{\vartheta_1 s_1^2 + \vartheta_2 s_2^2}{\vartheta_1 + \vartheta_2} = 5.42$$
$$\Rightarrow S = \sqrt{5.42} = 2.33$$

Test statistic:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{S\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} = \frac{12-11}{(2.33)\sqrt{\frac{1}{5}+\frac{1}{7}}} = 0.735$$
$$\Rightarrow |\gamma| = 0.735$$

Conclusion: $|\gamma| < t_2$. we accept the Null hypothesis

4.A Certain Stimulus administered to each of 12 patients resulted in the following increases of blood pressure:
5,2,8 -1,3,-2,1,5,0,4,6. Can it be concluded that the stimulus will in general be accompanied by an increase in blood pressure?

Solution: To test whether the mean increase in blood Pressure of all patients to whom the stimulus is administered Will be positive. We have to assume that this population is normal with mean ' μ ' and S.D. 'σ' which are unknown.

$$\left.\begin{array}{c}\text{Now the Null hypothesis is} \\ H_0: \mu = 0 \\ \text{Alternative hypo thesis, } H_1: \mu_1 > 0\end{array}\right\} \rightarrow (1)$$

The test statistic under null hypothesis H_0 is

$$t = \frac{\bar{d}}{\left(\frac{s}{\sqrt{n-1}}\right)} \sim t(n-1 \text{ degrees of freedom })$$

By the given data,

$$\bar{d} = \frac{5+2+8-1+3+0+6-2+1+5+0+4}{12} = 2.583$$

$$s^2 = \frac{\sum d^2}{n} - \bar{d}^2$$

$$= \frac{1}{12}[5^2 + 2^2 + 8^2 + (-1)^2 + 3^2 + 0^2 + 6^2 + (-2)^2 + 1^2 + 5^2 + 0^2 + 4^2] - (2 \cdot 583)^2$$

$$= 8.744$$

$$\Rightarrow s = \sqrt{8.744} = 2.9571$$

$$\therefore t = \frac{\bar{d}}{\left(\frac{8}{\sqrt{n-1}}\right)} = \frac{2.583}{\left(\frac{2.9571}{\sqrt{12-1}}\right)}$$

$$= \frac{(2.583)\sqrt{11}}{2.9571}$$

$$= 2.897 - t((n-1)d.f))$$

Conclusion: The tabulated value of $t_{0.05}$ at 11d. f is 2.2 .
$\therefore |t| > t_{0.05}, H_0$ is rejected.
that is, the stimulus does not increase the blood pressure. The stimulus in general will be accompanied by increase in blood pressure.

5) The following results are obtained from a sample of 10 boxes in biscuits. Mean weight is 490 gm , Standard deviation is 9 gm . Could the sample Come from a population having a mean of 500 gm ?

Sol: Given that,

$$n = 10$$
$$\bar{X} = 490 \text{ gm}, \sigma^2 = s = 9 \text{ gm}$$
$$\text{mean } \mu = 500$$

Now by the definition

$$s = \sqrt{\frac{n}{n-1}}\, s^2 = \sqrt{\frac{10}{9} \times 9^2} = \sqrt{90}$$

$$= 9.486$$

The Null hypothesis is,
H_0 : The difference is not significant

i.e., $H_0: \mu = 500$

Alternative hypothesis $H_1: \mu \neq 500$

$$\text{By t-test, } t = \frac{\overline{X} - \mu}{\left(\frac{s}{\sqrt{n}}\right)} = \frac{490 - 500}{\left(\frac{9.486}{\sqrt{10}}\right)}$$

$$= -0.333$$

By tabulated value $t_{0.05} = 2.26$ for $\gamma = 9$

conclusion: Since $|t| = 0.333 > t_{0.05}$ the hypothesis
H_0 is rejected i.e. $\mu \neq 500$

∴ The Sample Could not have come from the population having mean 500 gms.

13.9 EXERCISE

1)The average hourly wage of a sample of a Sample of 150 workers in a point 'A' was Rs 2.56, with a S.D. of Rs 1.08. The average hourly wage of a sample of 200 workers in plant B was Rs. 2.87 with a S. D. of Rs 1.28. can an applicant Safely assume that the hourly wages paid by plant B are higher than those paid by plant A?

2)Samples of Sizes 10 and 14 were taken from two normal populations with S.D. 3.5 and 5.2. The Sample means were found to be 20.3 and 18.6. Test whether the means of the two populations are the Same at 5% Level.

3)The heights of Six randomly chosen sailors are (in inches): 63, 65, 68, 69, 71 and 72. Those of 10 randomly Chosen soldiers are 61, 62, 65, 66, 69,69, 70, 71, 72 and 73. Discuss the light that these data thrown on the suggestions that Sailors are on the average taller than Soldiers.

(4)To test the claim that the resistance of electric wire can be reduced by at least 0.05 ohms by alloying 25 Values obtained for each alloyed wire and Standard wire produced the following results:

	Mean	S.D
Alloyed wire	0.083 ohms	0.003 ohms
Standard wire	0.136 ohms	0.002 ohms

Test at 5% level whether or not the claim is sustained.

5)A group of 5 patients treated with medicine A weigh 42, 39, 48, 60 and 41 kgs. Second group of 7 patients from the same hospital treated with medicine B weigh 38, 42, 56, 64, 68, 69 and 62 kgs. Do you agree with the claim that medicine B increases the weigh significantly?

6)To examine the hypothesis that the husbands are more intelligent than the wives, an investigator took a sample of 10 couples and administered them a test measures the I.Q. The results are as follows.

Husband	117	105	97	105	123	109	86	78	103	107
Wives	106	98	87	104	116	95	90	69	108	85

7) Examine the truth value of the hypothesis at level of significance of 0.05.
Two independent samples of 8 & 7 items respectively had the following values.

Sample I	11	11	13	11	15	9	12	14
Wives	9	11	10	13	9	8	10	

Is the difference between the means of samples significant?

13.10 SUMMARY

Comparing two populations is an important topic in statistics because it occurs quite frequently. Moreover, we can use inference to move beyond just looking at two sample means, as was suggested in our driving example. We can go on to figure out whether the difference between two groups is statistically significant; and if it is, we can calculate a confidence interval for the difference of population means.

We conclude this lesson with one final comment related to checking the underlying assumptions for the two-sample t-procedures. In the development of the two-sample t-procedures for cases where the sample sizes are small, we assume that the population distributions are normal. As it turns out, if the sample sizes are reasonably close and the population distributions are similar in shape, without major outliers, the probabilities from the t-distribution are quite accurate even if the population distributions are not normal.

13.12 TECHNICAL TERMS

- Independent Samples
- Dependent Samples (Paired Samples)
- Hypothesis Testing

- t-Test for a Single Mean
- t-Test for the Difference of Means
- Degrees of Freedom (df)
- P-Value

13.13 SELF ASSESSMENT QUESTIONS

SHORT:

1. What is the difference between independent and dependent samples?
2. When is a t-test for a single mean used?
3. What assumptions are required for using a t-test for two small samples?
4. How do degrees of freedom affect the t-test?
5. What does the p-value indicate in hypothesis testing?

ESSAY:

1. Explain the concept of hypothesis testing and its importance in statistical analysis.
2. Discuss the difference between independent and dependent samples with examples.
3. Describe the steps involved in conducting a t-test for a single mean.
4. Compare and contrast the t-test for a single mean and the t-test for the difference of means of two small samples.
5. Explain how sample size and variance affect the results of a t-test.

13.13 FURTHER READINGS

1. Introduction to Probability and statistics by J. Susan Milton and J.C. Arnold, 4ed, TMH(2007)

2. Mathematical Statistics by R.K. Goel, Krishna Prakasan Media (P) ltd., Meerut

3. Fundamentals of Mathematical Statistics by S.C. Gupta and V.K. Kapoor, Sultan Chand & sons, New Delhi.

Dr. T.V. Pradeep Kumar

LESSON-14

COMPARING TWO VARIANCES

OBJECTIVES

After Completion of This lesson the student should able to understand about

(i) Inferences about population variances

(ii) F-Distribution

(iii) Some real applications of F- distribution

STRUCTURE:

14.1 Introduction

14.2 Inferences about a population Variana

14.3 F-Distribution

14.4 Construct and interpret a confidence interval for two population variance.

14.5 Applications of f- distribution.

14.6 Relation between 't' and f distributions

14.7 Solved Problems

14.8 Exercise

14.9 Summary

14.10 Technical Terms

14.11 Self Assessment Questions

14.12 Further Reading

14.1 INTRODUCTION:

So far, we considered inference to Compare two proportions and inference to Compare two means. In this lesson, we will present how to Compare two population Variances.

why would we want to compare two population Variances? In many situations, such as in quality Control problems, where we may want to choose the process with smaller variability for a variable of interest.

one of the essential steps of a test to compare two population variances is for checking the equal variances assumption if we want to use the pooled variance.

Many people use this test as a guide to see if there are any clear violations, much like Using the rule of thumb.

14.2 INFERENCES ABOUT A POPULATION VARIANCE

Suppose we want to test if a random sample x_i , i=1,2,…,n has been drawn from a normal population with specified Variance $\sigma^2=\sigma_0^2$(Say)

Under the null hypothesis that the population Variance is $\sigma^2=\sigma_0^2$, the statistic χ^2 the statistic,

$$\chi 2=\sum_{i=1}^{n}\left[\frac{(x_i-\bar{x})^2}{\sigma_0^2}\right]$$

$$=\frac{1}{\sigma_0^2}\left[\sum_{i=1}^{n}x_i^2-\frac{(\sum x_i)^2}{n}\right]$$

$$=\frac{ns^2}{\sigma_0^2} \qquad \rightarrow (1)$$

which follows chi-square distribution with $(n-1)$ degrees of freedom.

Here by comparing the calculated value with the tabulated value of χ^2 for (n-1) d .f . at certain Level of Significance (usually 5%), we may retain or reject the null hypothesis.

14.2.1 Remarks

(1) (a) The above test can be applied only if the population from which the sample is drawn is normal

(2) If the sample size ' n ' is large ($n>30$), then we can use fisher's approximation

$$\sqrt{2\chi 2}\sim N(\sqrt{2n-1},1)$$

that is, $Z=\sqrt{2\chi 2}-\sqrt{2n-1}\sim N(0,1)$ and then apply Normal test.

14.2.2 Example: It is believed that the precision (as measured by the variance) of an instrument is no more than 0.16 , write down the null and alternative hypothesis for testing this brief. Carry out the test at 1% level given 11 measurements of the same subject on the instrument:

Measurements:2.5, 2..3, 2.4, 2.3, 2.5, 2.7, 2.5, 2.6, 2.6, 2.7, 2.5

Solution: from the given data of above,

Null hypothesis: $H_0: \sigma^2 = 0.16$

Alternative hypothesis: $H_1: \sigma^2 > 0.16$

a) under the null hypothesis,

$$H_0: \sigma^2 = 0.16, \text{the statics is}$$

$$\chi 2 = \frac{ns^2}{\sigma^2} = \sum \frac{(x - \bar{x})^2}{\sigma^2} \qquad \text{(See table in next page)}$$

$$\Rightarrow \bar{X} = 2.51, \sum (X - \bar{X})^2 = 0.1891$$

So $\chi 2 = \frac{0.1891}{0.16} = 1.182$

Which follows χ^2-distribution with $(n-1)$ d.f.

$$\Rightarrow (11 - 1) = 10 \text{ degrees as freedom.}$$

Computation of Sample variance as follows

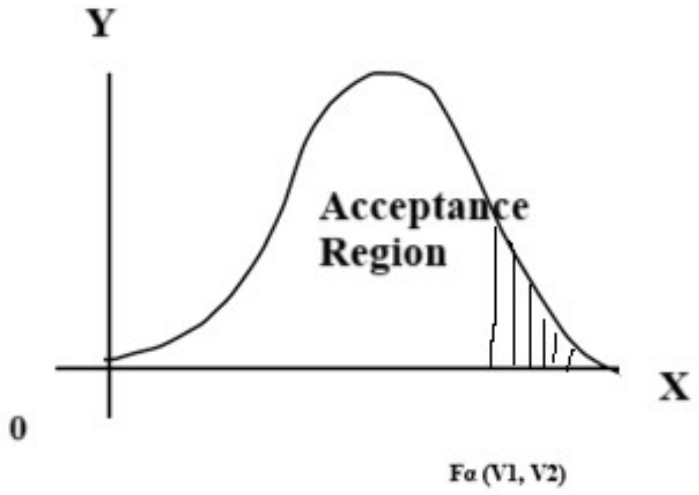
	X	$X - \bar{X}$	$(X - \bar{X})^2$
	2.5	-0.01	0.0001
	2.3	-0.21	0.0441
	2.4	-0.11	0.0121
	2.3	-0.21	0.0441
	2.5	-0.01	0.0001
	2.7	0.19	0.0361
	2.5	-0.01	0.0001
	2.6	0.09	0.0081
	2.6	0.09	0.0081
	2.7	0.19	0.0361
	2.5	-0.01	0.0001
Total	27 .6		0.1891

$$\bar{X} = \frac{27.6}{11} = 2.51$$

Conclusion: Since the calculated value of χ^2 is less than the tabulated value 23.2 of χ^2 for 10 degree of freedom at 1% level of Significance, it is not Significant. Hence H_0 may be accepted and we conclude that the data are consistent with the hypothesis that the precision of the instrument is 0.16

14.3 F-DISTRIBUTION

suppose we have a normal population with mean μ_1 and variance σ_1^2, and another normal postulation with mean μ_2 and variance σ_2^2.



Two samples are drawn from the different population.

First sample size is n_1 and its variance is say S_1^2, Second sample size is n_2 and its variance is S_2^2.

The Sampling distribution of the ratio of the Variances of the two independent random samples given by

$$F = \frac{\left(\frac{s_1^2}{\sigma_1^2}\right)}{\left(\frac{s_2^2}{\sigma_2^2}\right)} = \frac{s_1^2\sigma_2^2}{s_2^2\sigma_1^2} \quad \rightarrow \quad (1)$$

with $V_1 = n_1 - 1$ and $V_2 = n_2 - 1$ degrees of freedom is Known as F-distribution.

If the variances of the populations are same, that is, $\sigma_1^2 = \sigma_2^2$, then

$$F = \frac{s_1^2}{s_2^2} \quad \rightarrow \quad (2)$$

Generally Distribution curve lies in the first quadrant.
$F_\alpha(v_1, v_2)$ is the value of F with v_1, v_2 i.e, $n_1 - 1, n_2 - 1$ degrees of freedom such that area under the f-distribution curve to the right of F_α is equal to α.

Then $F_{1-\alpha}(v_1, v_2) = \frac{1}{F_\alpha(v_2, v_1)}$

Here $F_\alpha(v_1, v_2)$ value can be found by the F-distribution table.

for $\alpha = 0.05$ and $\alpha = 0.01$ are level of Significance.

If $v_1 = v_2$, then the probability density function of F and $\frac{1}{F}$ are the same.

14.3.1 Example: If two independent random samples of sizes $n_1 = 9$ and $n_2 = 16$ are taken from a normal population, what is the probability that the variance of the first Sample will be at least 4 times as large as the variance of the second sample.

Solution: Given that

$$n_1 = 9, \ n_2 = 16$$
$$v_1 = 9 - 1 = 8$$
$$v_2 = 16 - 1 = 15$$

from the f-table

$$F_{0.01}(v_1, v_2) = F_{0.01}(8,15) = 4.00 \quad \longrightarrow (1)$$

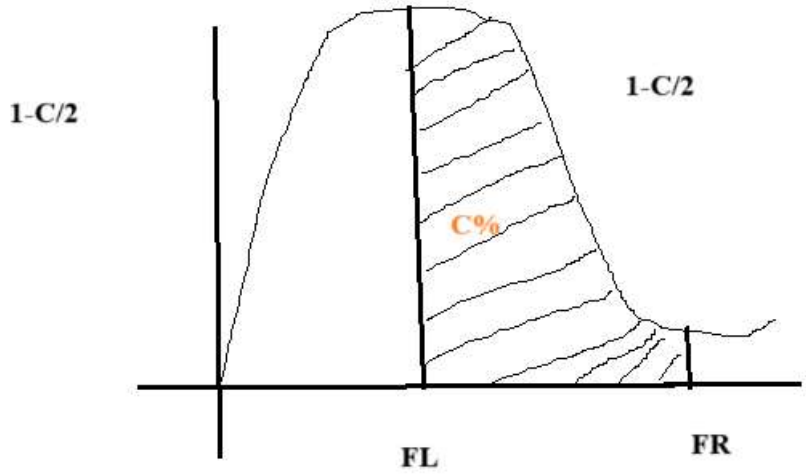
We have the Variance of the first Sample will be at least 4 times as large as variance of the second sample

$$F(v_1, v_2) = \frac{s_1^2}{s_2^2} = \frac{4s_2^2}{s_2^2} \text{ (by the given data)}$$

$$= 4$$
$$\therefore F(v_1, v_2) = 4 \qquad \rightarrow (2)$$

For (1) and (2), the desired probability is 0.01 .

14.4 CONSTRUCT AND INTERPRET A CONFIDENCE INTERVAL FOR TWO POPULATION VALANCES



Suppose a Sample of size n_1 with sample variance s_1^2 is taken from population 1 and a sample of size n_2 with sample variance s_2^2
is taken from population 2, where the populations are independent and normally distributed. For the confidence interval with Confidence level C for the ratio of the population Variances $\frac{\sigma_1^2}{\sigma_2^2}$ are

$$\text{Lower limit} = \frac{1}{F_L} \times \frac{s_1^2}{s_2^2}$$

$$\text{Upper limit} = \frac{1}{F_R} \times \frac{s_1^2}{s_2^2}$$

When F_L is the F-score so that the area in left tail of the F-distribution is $\frac{1-c}{2}$, F_R is the F-Score so that the area in the right tail of the F-distribution is $\frac{1-C}{2}$ and the F-distribution has $n_1 - 1$ and $n_2 - 1$ degrees of freedom.

14. 4.1 Note

(1) Like the other Confidence intervals, we have been, the F-Scores are the Values that trap C % of the observations in the middle of the distribution so that the area of each tail is $\frac{1-c}{2}$.

(2) Since F-distribution is not Symmetric, the Confidence interval for the ratio of the population Variances requires that we calculate two different F-Scores.

i.e one for the left tail and one for the right tail.

(3) It is important that the populations are Independent and normally distributed.

If the populations are not normal then the confidence interval will not give an accurate result.

14.2 Example: Two local walk in medical clinics want to determine if these is any Variability is the time Patients wait to see a doctor at each clinic. In a sample of 30 patients at clinic 1 , the standard deviation for the wait time be see a doctor was 45 minutes. In a sample of 40 patients at clinic 2, the standard deviation for the wait time to see a doctor was 27 minutes. Assume the population of Wait time at the two clinics are independent and normally distributed.

a) Construct 95% Confidence interval for the ratio of the variances for the wait times at the two clinics.

b) Interpret the Confidence interval found in (a)

c) Is there evidence to suggest that there is a difference in the variances of the wait times at the two clinics ? Explain

Solution :(a) Let us consider

Clinic-1 be population-1

and clinic -2 be prpulation-2

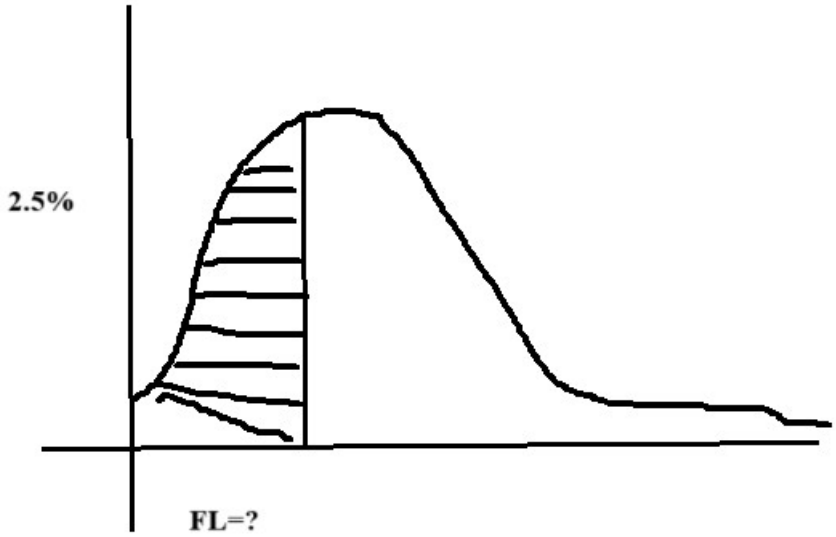
from the question, given above, we have to following information

Clinic-1	Clinic-2
$n_1=30$	$n_2=40$
$s_1^2=45^2=2025$	$s_2^2=27^2=729$

Now we have to find the confidence interval, To find this we need to find the F_L Scorns tor the 95% confidence interval.

Tho means that we have to find F_L score So that the area in the left tail is

$$\frac{1-0.95}{2}=0.025$$

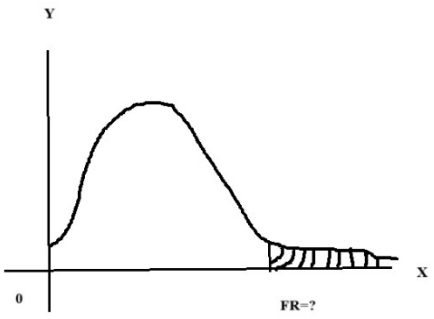


Function	f.1nv	Answer
field-1	0.025	0.4919
field-2	29	
field-3	39	

Also, we have to find the F_R - score for the 95% Confidence interval. This means that we need to find F_R-Score so that the area in the right tail is $\frac{1-0.95}{2} = 0.025$.
Now the degree of freedom for $F-$ distributed are

$$d.f(F_1) = n_1 - 1 = 30 - 1 = 29$$

$$d \cdot f(F_2) = n_2 - 1 = 40 - 1 = 39$$



Function	f inv rt	Answer
field -1	0.025	1.9618
Field -2	29	-
Field- 3	39	-

So that,

$$F_L = 0.4919, \text{and } F_R = 1.9618$$

Now 95% Confidence interval is

$$\begin{aligned}\text{Lower limit} &= \frac{1}{F_R} \times \frac{s_1^2}{s_2^2} \\ &= \frac{1}{1.9618} \times \frac{2025}{729} \\ &= 1.416 \\ \text{Upper limit} &= \frac{1}{F_L} \times \frac{s_1^2}{s_2^2} \\ &= \frac{1}{0.4919} \times \frac{2025}{729} \\ &= 5.646\end{aligned}$$

Confidence interval is

$$\text{I} = (1.416, 5.646)$$

b) We are 95% Confidence that the ratio of the variances in the wait times at the two clinics is between 1.416 and 5.646 .

(c) Because, 1 is outside the confidence interval, it Suggests that the ratio of the variances is $\frac{\sigma_1^2}{\sigma_2^2}$ is not 1 .

If the ratio of the variances cannot equal to 1 , then the variances cannot be equal. So there is a difference in the variances of the wait times at the two clinics.

14.5 APPLICATIONS OF F -DISTRIBUTION

F-distribution has the following some applications in statistical theory Under H_0

(1) Testing the equality of variances of two normal populations.
(2) Testing the equality of means of $k(> 2)$ normal populations
(3) Carrying out analysis of variance for two-way classified date.

14.6 RELATION BETWEEN *t* AND *F* DISTRIBUTIONS

In F-distribution with $(\vartheta_1, \vartheta_2)$ degas of freedom, tale $\vartheta_1 = 1, \vartheta_2 = v$ and $t^2 = F$, that is, $dF = 2tdt$

So, the probability differential of F transform to,

$$dG(t) = \frac{(1/\vartheta)^{\frac{1}{2}}}{\beta\left(\frac{1}{2}, \frac{\vartheta}{2}\right)} \cdot \frac{(t^2)^{\frac{1}{2}-1}}{\left(1 + \frac{t^2}{\vartheta}\right)^{\frac{(\vartheta+1)}{2}}} 2tdt, \quad 0 \leq t^2 < \infty$$

$$= \frac{1}{\sqrt{\vartheta} \quad \beta\left(\frac{1}{2}, \frac{\vartheta}{2}\right)} \cdot \frac{1}{\left(1 + \frac{t^2}{\vartheta}\right)^{\frac{\vartheta+1}{2}}} dt \rightarrow (1)$$

The factor ' 2 ' disappearing the total probability in the range $(-\infty, \infty)$ is unity.

This is the probability function of t-distribution with ' ϑ ' degrees of freedom.

Hence we have the following relation between 't' and 'F' distributions:

If a statistics ' t ' follows students ' t ' distribution with ' n ' degrees of freedom.

Then t^2 follows F-distribution with $(1, n)$ degrees of freedom.
we show t symbolically as.

If $t \sim t_{(n)}$ then $t^2 \sim \mathrm{F}_{(1,n)} \quad \rightarrow (2)$

14.6.1 Example. Two Random samples gave the following results

Sample	Size	Sumple mean	Sum of squares of deviations from the mean
1	10	15	
2	12	14	

Test whether the sample Come from the same normal population at 5%. Level of significance

($F_{0.05}(9, 11) = 2.90, F_{0.05}(11, 9) = 3.10$ appoxi and $t_{.0.05}(20) = 2.086$ and $t_{0.05}(22) = 2.07$)

Solution: A normal population has two parameters, say Mean μ' and variance σ^2.

To test it two independent samples have been drawn from the same normal population, we have to test

(i) the equality of populate on means,
(ii) the equality of population variances.

Now Null hypotheses
The two samples have been drawn from the same normal population.

i.e., $H_0: \mu_1 = \mu_2$ and $\sigma^2{}_1 = \sigma_2^2 \quad \rightarrow (1)$

Then we have that Equality of means will be tested by applying ' t '-test and equality of variances will be tested by applying F-test.

Since t-test assumes $\sigma_1^2 = \sigma_2^2$ we shall first apply f-test, and then t-test.

from the given data,n_2

$$n_1 = 10, \quad \overline{x_1} = 15 \qquad n_2 = 12, \ \overline{x_2} = 14$$

$$= \sum (x_1 - \overline{x_1})^2 = 90$$
$$\sum (x_2 - \overline{x_2})^2 = 108$$

F- test

$$\text{since } S_1^2 = \frac{1}{n_1 - 1} \sum (x_1 - x_2)^2$$

$$= \frac{90}{9} = 10$$
$$S_2^2 = \frac{1}{n_2 - 1} \sum (x_2 - \bar{x}_2)^2$$
$$= \frac{108}{11} = 9.82$$

Clearly, $S_1^2 > S_2^2$,
under $H_0: \sigma_1^2 = \sigma_2^2$
then the test statistic

$$F = \frac{S_1^2}{S_2^2} \sim F\big((n_1 - 1), (n_2 - 1)\big) = F(9,11)$$
$$\Rightarrow F = \frac{S_1^2}{S_2^2} = \frac{10}{9.82} = 1.018$$

From f-distribution table

$$F_{0.05}(9,11) = 2.90$$

since calculated F is less than tabulated value..
⇒ It is not significant.
Hence null hypothesis of equality of population variances may be accepted.

Since $\sigma_1^2 = \sigma_2^2$ we can now apply ' t ' test for testing $H_0: \mu_1 = \mu_2$

t-test : Under $H_0^1 = \mu_2$, against alternative hypothesis , $H_1^1: \mu_1 \neq \mu_2$,

The test statistic is,

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim t_{n_1+n_2-2} = t_{20}$$

Where $s^2 = \frac{1}{n_1+n_2-2}\left[\sum (x_1 - \bar{x}_1)^2 + \Sigma(x_2 - \bar{x}_2)^2\right]$

$$= \frac{1}{20}(90 + 108) = 9.9$$

$$t = \frac{15 - 14}{\sqrt{9.9\left(\frac{1}{10} + \frac{1}{12}\right)}} = \frac{1}{\sqrt{9.9 \times \frac{1}{60}}} = \frac{1}{\sqrt{1.815}}$$
$$= 0.742.$$

Tabulated $t_{0.05}$ for 20 d.f =2.086

Since$|t| < t_{0.05,}$ it is not significant.

Hence the hypothesis $H_0^1: \mu_1 = \mu_2$ may be accepted.

Since both hypothesis $H_0^1: \mu_1 = \mu_2$ and$H_0: \sigma_1^2 = \sigma_2^2$ are accepted.

We may regard that given samples have drawn from the same normal population.

been drawn from the same normal population.

14.6.2 Example: The two random Samples reveal the following data

Sample	**Size(n)**	**Mean(μ)**	**Variance(σ^2)**
I	**16**	**440**	**40**
II	**25**	**460**	**42**

Test whether the samples come from the same normal population

Solution: A normal population has two parameters namely the mean μ and the variance σ^2. To test whether the two independent Samples have been drawn from the same normal population, we have to test

(i) the equality of means; (ii) the equality of Variances

Since the t-test assumes that the sample variances are equal, we first apply F-test.

F-test:

Null hypothesis :$H_0: \sigma_1^2 = \sigma_2^2$

Clearly the population Variance do not differ Significantly.

Alternative Hypothesis:$H_1: \sigma_1^2 \neq \sigma_2^2$

Under the null hypothesis, the test statistic is given by

$$F = \frac{s_1^2}{s_2^2}, (s_1^2 > s_2^2)$$

Given that n_1=16, n_2 =25, s_1^2=40, s_1^2=42

$$F = \frac{s_1^2}{s_2^2} = \frac{n1s_1^2}{n-1} / \frac{n2s_2^2}{n-1} = \frac{16\times40}{15} \times \frac{24}{25\times42} = 0.9752$$

Conclusion: The Calculated value of F is 0.9752.

The tabulated value $F_{0.05}$ (15,24)=2.11 Since the Calculated value is less than that of the tabulated value, H_0 is accepted,

i.e. the population Variances are equal.

t-test

Null hypothesis.
$H_0: \mu_1 = \mu_2$, i.e, the population means are equal.

Alternative hypothesis: $H_1: \mu_1 \neq \mu_2$

Under the null hypothesis the test statistic.

Given $n_1 = 16, n_2 = 25, \bar{X}_1 = 440, \overline{X_2} = 460$

$$s^2 = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2} = \frac{16 \times 40 + 25 \times 42}{16 + 25 - 2}$$
$$= 43.33$$
$$\therefore s = \sqrt{43.33} = 6.582$$
$$\Rightarrow t = \frac{\bar{X}_1 - \bar{X}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{440 - 460}{6.582\sqrt{\frac{1}{16} + \frac{1}{25}}}$$
$$= -9.490$$

for $(n_1 + n_2 - 2)$ degrees of freedom.

Conclusion: The calculated value of |t| is 9.490 .

The tabulated value of 't' at 36 d .f. for 5%. Level of significance is 1.96

Since the Calculated value is greater than tabulated value, so H_0 is rejected. This shows there is significant difference between means; that is $= \mu_1 \neq \mu_2$

Note: on confidence limits

NOTES

1. When calculating the limits for the confidence interval keep all of the decimals in the FF-scores and other values throughout the calculation. This will ensure that there is no round-off error in the answer. You can use Excel to do the calculations of the limits, clicking on the cells containing the FF-scores and any other values.

2. When writing down the interpretation of the confidence interval, make sure to include the confidence level and the actual ratio of population variances captured by the confidence interval (i.e. be specific to the context of the question). In this case, there are no units for the limits because variance does not have any limits.

Steps to Conduct a Hypothesis Test for Two Population Variances

1. Write down the null hypothesis that there is no difference in the population variances:

$$H_0: \sigma_1^2 = \sigma_2^2$$

The null hypothesis is always the claim that the two population variances are equal

2. Write down the alternative hypotheses in terms of the difference in population variances. The alternative hypothesis will be one of the following

$H_a: \sigma_1^2 < \sigma_2^2$

$H_a: \sigma_1^2 > \sigma_2^2$

$H_a: \sigma_1^2 \neq \sigma_2^2$

3. Use the form of the alternative hypothesis to determine if the test is left-tailed, right-tailed, or two-tailed.

4. Collect the sample information for the test and identify the significance level α .

5. Use the F-distribution to find the p-value (the area in the corresponding tail) for the test. The F-score and degrees of freedom are

$$F = \frac{s_1^2}{s_2^2}$$

$df_1 = n_1 - 1$

$df_2 = n_2 - 1$

6. Compare the p-value to the significance level and state the outcome of the test:

If p-value $\leq \alpha$, a reject H_0 in favour of H_a.

* The results of the sample data are significant. There is sufficient evidence to conclude that the null hypothesis H_0 is an incorrect belief and that the alternative hypothesis H_a most likely correct.

If p-value > α. do not reject H_0.

* The results of the sample data are not significant. There is not sufficient evidence to conclude that the alternative hypothesis H_a may be correct.

7 Write down a concluding sentence specific to the context of the question.

14.7 SOLVED PROBLEMS

Problems

(1) In a sample of 8 observations, the sum of squared deviations of items from the mean was 84.4 . In another example of 10 observations, the value was found to be 102.6 test whether the difference is significant at 5% level.

Solutions:

According to the data
(Σ) (2) (-) x $\bar{x}$

sum of squared deviations of items from the mean $\Rightarrow \boldsymbol{\Sigma}(x-\bar{x})^2$

$$\text{so, we get, } \boldsymbol{\Sigma}(x-\bar{x})^2 = 84.4 \text{ and } \Sigma(y-\bar{y})^2 = 102.6$$
$$\text{also } n_1 = 8 n_2 = 10.$$
$$S_1^2 = \frac{\sum (x-\bar{x})^2}{n_1 - 1} = \frac{84.4}{7} = 12.06$$
$$S_2^2 = \frac{\sum (y-\bar{y})^2}{n_2 - 1} = \frac{102.6}{9} = 11.4$$
$$F = \frac{\text{large er var iance}}{\text{smaller variance}} = \frac{S_1^2}{S_2^2} = \frac{12.06}{11.4} = 1.06$$

Null hypothesis $(\mathbf{H_0})$: $H_0: \sigma_1 = \sigma_2$

Alternative hypothesis $(\mathbf{H_1}): \sigma_1 \neq \sigma_2$

v_1 = degree of freedom for sample having large variance
= degree of freedom w.r.t to $s_1^2 = 8 - 1 = 7$
v_2 = degree of freedom for sample having smaller variance
= degree of freedom w.r.t to $s_2^2 = 10 - 1 = 9$
Given LOS $\alpha = 5\% = 0.05$

F -table value: $F_\alpha(\vartheta_1, \vartheta_2) = F_{0.05}(7,9) = 3.293.$

The calculated value of F is less than the table value. Hence, we accept the hypothesis.

2. Time taken by workers in performing a job by method 1 and method 2 is given below.

A: 66 67 75 76 82 84 88 90 92

B: 64 66 74 78 82 85 87 92 93 95 97

Does the data show that variance of time distribution from population with these samples are drawn do not differ significantly at 1% level?

Solution:

X	$\bar{X}$	$X-\bar{X}$	$(X-\bar{X})^2$	Y	$\bar{Y}$	$Y-\bar{Y}$	$(Y-\bar{Y})^2$
66		-14	196	64		-19	361
67		-13	168	66		-17	289
75		-5	25	74		-9	81
76		-4	16	78		-5	25
82		+2	4	82		-1	1
84	$\bar{X}=\frac{\sum X}{n_1}$	+4	16	85	$\bar{Y}=\frac{\sum Y}{n_2}$	+2	4
88	$=\frac{720}{9}$	+8	64	87	$=\frac{913}{11}$	+4	16
90	=80	+10	100	92	=83	+9	81
92		+12	144	93		+10	100
				95		+12	144
				97		+14	196
$\sum X=$ **720**			$\sum(X-\bar{X})^2$=**734**	$\sum Y$=**913**			$\sum(Y-\bar{Y})^2$ **= 1298**

$$S_1^2=\frac{\sum\ (X-\bar{X})^2}{n_1-1}=\frac{734}{9-1}=91.75$$

$$S_2^2=\frac{\sum\ (y-\bar{y})^2}{n_2-1}=\frac{129.8}{11-1}=129.8$$

Calculated value

$$F=\frac{larg\ er\ variance}{smaller\ variance}=\frac{{S_2}^2}{{S_1}^2}=\frac{129.8}{91.75}=1.415$$

Null hypothesis ($\mathbf{H_0}$): $H_0: \sigma_1=\sigma_2$

Alternative hypothesis ($\mathbf{H_1}$): $\sigma_1 \neq \sigma_2$

$\mathbf{v}_1$ = degree of freedom for sample having large variance
= degree of freedom w.r.t to $s_2^2 = 11 - 1 = 10$
v_2 = degree of freedom for sample having smaller variance
= degree of freedom w.r.t to $s_1{}^2 = 9 - 1 = 8$
Given LOS $\alpha = 1\% = 0.05$

F -table value: $F_\alpha(\vartheta_1, \vartheta_2) = F_{0.01}(10,8) = 5.814$.

Since F $< F_\alpha$, we accept the null hypothesis.

3. Pumpkins were grown under two experimental conditions. Two random samples of 11 and 9 pumpkins. The sample standard deviation of their weights as 0.8 and 0.5 respectively. Assuming that the weight distributions are normal, Examine the truth value of hypothesis that the true variances are equal.

Sol: According to the data

we get, $S_1 = 0.8$ and $S_2 = 0.5$

also $n_1 = 11, n_2 = 9$.

$S_1^2 = 0.8^2 = 0.64$
$S_2^2 = 0.5^2 = 0.25$
$$F = \frac{\text{larger variance}}{\text{smaller variance}} = \frac{S_1^2}{S_2^2} = \frac{0.64}{0.25} = 2.56$$
Null hypothesis ($\mathbf{H_0}$): $H_0: \sigma_1 = \sigma_2$
Alternative hypothesis ($\mathbf{H_1}$): $\sigma_1 \neq \sigma_2$
v_1 = degree of freedom for sample having large variance
= degree of freedom w. r. t to $s_1{}^2 = 11 - 1 = 10$
v_2 = degree of freedom for sample having smaller variance
degree of freedom w.r.t to $s_2{}^2 = 9 - 1 = 8$
Take LOS $\alpha = 5\% = 0.05$
F table value: $F_\alpha(\vartheta_1, \vartheta_2) = F_{0.05}(10,8) = 3.347$.

The calculated value of F is less than the table value. Hence, we accept the hypothesis.

14.8 EXERCISE

1. The following random samples are measurements of the heat-producing capacity (in millions of calories per ton) of specimens of coal from two mines:

Mine 1	8,260	8,130	8,350	8,070	8,340	...
Mine 2	7,950	1,890	7,900	8,140	7,920	7,840

Use the 0.05 level of significance to Examine whether it is reasonable to assume that the variances of the two populations are equal.

2. In one sample of 10 observations, the sum of the deviations of the sample values from sample mean was 120 and in the other sample of 12 observations it was 314. Examine whether the difference is significant at 5% level.

3.The measurements of the output of two units have given the following results. Assuming that both samples have been obtained from the normal populations at 1% significant level, examine whether the two populations have the same variance.

4. Find the value of $F_{0.05}$(Corresponding to a left hand tail probability of 0.05) for v_1=10 and v_2=24 degrees of freedom (Ans:0.3649)

5. For an F-distribution find

a)$F_{0.05}$ with ϑ_1=6 and ϑ_2=15

b)$F_{0.01}$ with ϑ_1=24 and ϑ_2=30

(Ans:2.79, 2.47)

14.9 SUMMARY

In this Lesson, we discussed how to compare population parameters from two samples. It is important to recognize which parameters are of interest. Once we identify the parameters, there are different approaches based on what we can assume about the samples. We compared two population proportions for independent samples by developing the confidence interval and the hypothesis test for the difference between the two population proportions.

Next, we discussed how to compare two population Variances. The approach for inference is different if the samples are paired or independent. For two independent samples, we presented two cases based on whether or not we can assume the population variances are the same.

Finally, we discussed a test for comparing two sample variances from independent samples using the F-test.

In this Lesson, we considered the cases where the response is either qualitative or quantitative, and the explanatory variable is qualitative (categorical). In the next

Lesson, we will present the case where both the response and the explanatory variable are qualitative.

14.10 TECHNICAL TERMS

- Population Variance (σ2\sigma^2σ2)
- F-Distribution
- Confidence Interval for Variance Ratio
- F-Test
- Analysis of Variance (ANOVA)
- Relationship Between t and F Distributions
- Critical Value of F

14.11 SELF ASSESSMENT QUESTIONS

SHORT:

- What is the purpose of the F-distribution in statistical analysis?
- How is an F-test used to compare two population variances?
- What is the relationship between the t-distribution and the F-distribution?
- How do you construct a confidence interval for the ratio of two population variances?
- What are some common applications of the F-distribution?

ESSAY:

- Explain the concept of the F-distribution and its significance in comparing population variances.
- Describe the step-by-step procedure for conducting an F-test for equality of two variances.
- Discuss the construction and interpretation of a confidence interval for the ratio of two population variances.
- Compare and contrast the t-distribution and the F-distribution, explaining their relationship.
- Explain the applications of the F-distribution in real-world scenarios, including ANOVA and variance comparison tests.

14.12 Further Reading

1. Introduction to Probability and statistics by J. Susan Milton and J.C. Arnold, 4ed, TMH(2007)

2. Mathematical Statistics by R.K. Goel, Krishna Prakasan Media (P) Itd., Meerut

3. Fundamentals of Mathematical Statistics by S.C. Gupta and V.K. Kapoor, Sultan Chand & sons, New Delhi.

Dr. T.V. Pradeep Kumar

Lesson 15

ANALYSIS OF VARIANCE

OBJECTIVES:

Main aim of this lesson is to know for Factors and levels, multifactor design by ANOVA.

Learning out comes: After completion of this Lesson the Student Should able to understand

(i) one-way classification fixed effects model

(ii) Comparing Variances

(iii) Pair wise comparisons of ANOVA

STRUCTURE:

15.1 Introduction
15.2 Definition and examples
15.3 One way classification
15.4 ANOVA for fixed effect model
15.5 ANOVA- Two way classification
15.6 Assumptions for Analysis of Variance test
15.7 Solved problems.
15.8 Summary
15.9 Technical Terms
15.10 Self-Assessment Questions
15.11 Further Reading

15.1 INTRODUCTION:

The Analysis of variance (ANOVA) is a Very important statistical tool for test of Significance. The term Analysis of Variance (ANOVA) was introduced by a famous statistician Prof. R.A. Fisher for Separation of the experimentally observed variance into a number of Components traceable to specific sources.

The main aim of analysis of variance (ANOVA) is to find how much of the total variability is due to each factor and by comparing these Contributory amounts of Variation.

15.2. DEFINITION AND EXAMPLES

15.2.1 Definition(R.A Fisher): (Analysis of variance(ANOVA) is the separation of variance assignable to one group of causes from the variance assignable to other group.

15.2.2

Note: The first step in the ANOVA is to separate the total variation in the whole number of observations in to two parts.

(i) The variance between the classes

(ii)The variance which arises from individual differences with the classes.

15.2.3 Example

A group of cows may be divided into several classes according to their breeds and the amount of milk yield of each cow over a given period may regarded.

Then, we may examine the variation between the mean-milk-yield of different breeds and the variation with in breeds.

So, here the criteria of classification is only one, that is the breed of the cows.

15.2.4 Assumptions:

(i) The observations are independent.

(ii)Parent population from which observation are taken is normal

(iii)Various treatment and environment effects are additive in nature.

15.3 ONE WAY CLASSIFICATION:

In general, one way ANOVA techniques Can be used to study the effect of k (≥ 2) levels of single factor.

Suppose a sample of N values of a Variate 'x' is subdivided into k classes according to Some criterion of classification.

Let the i^{th} class consist of n_i members and let the j^{th} member of j^{th} class be denoted by x_{ij}.

Then $\sum_{i=1}^{k} n_i = N$

15.3.1

for example table 15.1

Class	Sample Observations	Total	Mean
1	x_{11} x_{12} ..... x_1n_1	T_1	$\overline{x_1}$

2	x_{21} x_{22} ..... x_1n_2	T_2	$\overline{x_2}$
⋮		⋮	⋮
i	x_{i1} x_{i2}.....x_in_i	T_i	$\overline{x_i}$
⋮		⋮	⋮
k	x_{k1} x_{k2}..... x_kn_k	T_k	$\overline{x_k}$

One way classified data

The total variation in the observation x_{ij} can be split into the following two components

(i) The variation between the classes or the variation due to different bases of classification, commonly known as treatments.

(ii) The variation with in the classes, i.e, the inherent variation of the random variable within the observations of a class.

15.3.2. Note: The main object of analysis of variance technique is to examine if there is significant difference between the class means in view of the inherent variability with in the Separate classes.

15.3.3 Example: Let us consider the effect of k different rations on the yield in milk of N cow's (of the same breed and stock) divided into "k' classes of sizes n_1, n_2,… n_k respectively.

Here $N=\sum_{i=1}^{n} n_i \quad \rightarrow(1)$

The Sources of Variations are

(i) Effect of the ration(treatment): t_i, i=1,..k.

(ii) Error (E) produced by numerous causes of such magnitude that they are not detected identified with the knowledge that we have and they together produce a variation of of random nature obeying normal low of error's.

Mathematical model:

In this case the linear mathematical model will be:

$x_{ij}=\mu_i+\in_{ij}=\mu+(\mu_i+\mu)+\in_{ij} \quad \rightarrow(1)$

$=\mu+\alpha_i+\in_{ij}$, i=1,2,…,k

J=1,2,…,n

(i) x_{ij} is the yield from the j[th] Cow, (j = 1, 2 ,…,n) fed on the i[th] ration (i = 1, 2 ,..,k) →(2)

(ii)μ is the general mean effect given by $\mu = \sum_{i=1}^{k} \frac{n_i \mu_i}{N}$ →(3)

Where μ_i is the fixed effect due to the i[th] ration. That is, If there were no treatment differences and no chance causes then the yield of each cow will be μ.

(iii) 'α_i' is the effect of the i[th] ration given by

$\alpha_i = \mu_i - \mu$, i=1,2,..k →(4)

i.e, the i[th] ration increases (or decreases) the yield by an amount α_i.

By (3) and (4) , we get

$$\sum_{i=1}^{k} n_i d_i = \sum_{i=1}^{k} n_i (\mu_i - \mu)$$

$$= \sum_{i=1}^{k} n_i \mu_i - \mu \sum_{i=1}^{k} n_i$$

=N. μ - μ .N=0 →(5)

(iv) $\in_{ij}$ is the error effect due to chance.

15.3.5 Example: In one way classification, if the mean Sum of squares between groups is found to be less than that within groups, would you infer that there are no significant difference among the groups, or would you suspect a further source of variation?

Solution: Let $S_t^2 = \frac{1}{k-1} \sum_{i=1}^{k} n_i \ (\overline{x_i} - \overline{\overline{x}})^2$ →(1)

$$S_p^2 = \frac{1}{n-k} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \overline{\overline{x}})^2 \quad \rightarrow(2)$$

where s_t^2 is the mean sum of Squares between groups, S_p^2 is the mean sum of squares within the groups, then if $s_t^2 \le s_p^2$ that is F≤1, the variance ratio.

Then the dispersion of the Sample means about the general mean is not greater than expected ordinarily from a set of k- samples taken from the Same population, then it is taken to be insignificant and no further test is necessary.

However if $s_t^2 > s_p^2$, that is, F > 1 then we will apply a variance ratio test of significance.

15.4 ANOVA FOR FIXED EFFECT MODEL:

The fixed effect or parametric model is used as

$x_{ij} = \mu_i + \in_{ij} \quad \rightarrow (1)$

$= \mu + \alpha_i + \in_{ij}, \quad i=1,2,\ldots,k$

$J=1,2,\ldots,n$

Where $\alpha_i = \mu_i - \mu$, x_{ij} is the yield from the i^{th} row and j^{th} column.

Here a) μ is the general mean effect given by $\mu = \sum_{i=1}^{k} \frac{n_i \mu_i}{N}$

b) α_i, i=1,2,…k due the i^{th} iteration given by $\alpha_i = \mu_i - \mu$, where $\sum_{i=1}^{k} \sum_{j=1}^{n}$,

(c) $\in_j$, j=1,2,…n due to the j^{th} variety (breed of cow) given by $\in_j = \mu_j - \mu$

Where $\mu_j = \frac{1}{k} \sum_{i=1}^{k} \mu_{ij}$, j=1,2,…n

15.5 ANOVA TWO-WAY CLASSIFICATION

Now we proceed to the case of two way classification. We classify our sample of N values of 'x' according to some quality A into K classes and according to another quality B into 'n' classes, so that N=n.k

Let the sample variable value in the i^{th} A-class and j^{th} B-class be x_{ij}

Let $\overline{x_j}$ be the mean of j^{th} row and $\overline{x_i}$ be the mean of i^{th} column.

Then, we have the following identity:

$$\sum_{i=1}^{k} \sum_{j=1}^{n} (x_{ij} - \overline{x})^2 = \sum_{i=1}^{k} \sum_{j=1}^{n} [(x_{ij} - \overline{x_i} - \overline{x_j} + \overline{x}) + (\overline{x_i} - \overline{x}) + (\overline{x_j} - \overline{x})]^2$$

$$= \sum_{i=1}^{k} \sum_{j=1}^{n} (x_{ij} - \overline{x_i} - \overline{x_j} + \overline{x})^2 + \sum_{i=1}^{k} \sum_{j=1}^{n} (\overline{x_j} - \overline{x})^2 + \sum_{i=1}^{k} \sum_{j=1}^{n} (\overline{x_i} - \overline{x})^2$$

$$= \sum_{i=1}^{k} \sum_{j=1}^{n} (x_{ij} - \overline{x_i} - \overline{x_j} + \overline{x})^2 + \sum_{i=1}^{k} n(\overline{x_i} - \overline{x})^2 + \sum_{j=1}^{n} k(\overline{x_j} - \overline{x})^2 \quad \rightarrow (1)$$

clearly product terms in the expansion vanish as in the case of one way classification.

It can be shown that if all the data are drawn from the same population with Variance σ^2, then three terms of the R.H.S of (1) are the independent estimates of (k-1)(n-1)σ^2, (k-1)σ^2 and (n-1)σ^2 respectively.

ANOVA for two way classification

Source of variation	Sum of Squares	Degrees of freedom	Existance of variance	Remarks
Between A classes	$\sum_{i=1}^{k} n(\bar{x}_i - \bar{x})^2$	k-1	$\frac{1}{n}\sum n(\bar{x}_i - \bar{x})^2 = Q_A$	(a)Test $\frac{Q_A}{Q_{AB}}$ for (k-1) and (k-1)(n-1) degrees of freedom using F-Table
Between B classes	$\sum_{j=1}^{n} k(\bar{x}_j - \bar{x})^2$	n-1	$\frac{1}{n-1}\sum k(\bar{x}_j - \bar{x})^2 = Q_B$	
Residual	$\sum_{i=1}^{k}\sum_{j=1}^{n}(x_{ij} - \bar{x}_i - \bar{x}_j + \bar{x})^2$	(k-1)(n-1)	$\frac{\sum_{i=1}^{k}\sum_{j=1}^{n}(x_{ij} - \bar{x}_i - \bar{x}_j + \bar{x})^2}{(k-1)(n-1)} = Q_{AB}$	
Total	$\sum_{i=1}^{k}\sum_{j=1}^{n}(x_{ij} - \bar{x})^2$	nk-1		(b)Test $\frac{Q_B}{Q_{AB}}$ for (n-1) and (k-1)(n-1) degrees of freedom using F-table

15.5.1 Example:

On a feeding experiment a farmer has four types of hogs denoted by I, II, III and IV. These types are each divided into three groups, which are fed varietal rations A, B, and C. Then the following results are obtained, the as number in the table being the gain in weight is pounds in the various groups.

	I	II	III	IV	Total
A	7.0	16.0	10.5	13.5	47.0
B	14.0	15.5	15.0	21.0	65.5
C	8.5	16.5	9.5	13.5	48.0
Total	29.5	48.0	35.0	48.0	160.5

Solution: The above computations yield the following results.

Sum of squares		df	Unbiased estimates
Rations	54.1250	2	27.06
Types	87.7292	3	29.24
Residual	28.2083	6	4.70

The test of significance of the variation in relations we refer

$$F=\frac{27.06}{4.70}=5.76$$

To the Snedecor's table where corresponding to (2,6) degrees of freedom,

We find 5.14 for the 5% point and 10.92 for 1% point.

Similarly, the test of significance of the variation between types we compute

$$F=\frac{20.24}{4.70}=6.2$$

The entries in the table for (3.6) degrees of freedom are 4.76 for 5% point and 9.78 for the 1% point.

Next the entries in the table for(3,6) degrees of freedom are 4.70% for 5% point and 9.78 for 1% form.

Now we conclude that there is a significant difference between breeds and between variates of rations at 5% point, but that neither is significant at the 1% point.

15.5.2 Example:

Estimate the parameters of the one way classification model for the tin-coating weights given in the following table.

	Lab A	Lab B	Lab C	Lab D
	0.25	0.18	0.19	0.23
	0.27	0.28	0.25	0.30
	0.22	0.21	0.27	0.28
	0.30	0.23	0.24	0.28
	0.27	0.25	0.18	0.24
	0.28	0.20	0.26	0.34
	0.32	0.27	0.28	0.20
	0.24	0.19	0.24	0.18
	0.31	0.24	0.25	0.24
	0.26	0.22	0.20	0.28
	0.21	0.29	0.21	0.22

	0.28	0.16	0.19	0.21
TOTAL	3.21	2.72	2.76	3.00

Solution: Here N=12+12+12+12=48

Grand total for four labs are

T=3.21+2.72+2.76+3.00=11.69

Now $\mu = \frac{11.69}{48} = 0.244$

$\overline{x_1} = \frac{3.21}{12} - 0.244 = 0.024$

$\overline{x_2} = \frac{2.72}{12} - 0.244 = -0.017$

$\overline{x_3} = \frac{2.76}{12} - 0.244 = -0.014$

And $\overline{x_4} = \frac{3.00}{12} - 0.244 = 0.006$.

15.6 ASSUMPTIONS FOR ANALYSIS OF VARIANCE TEST

(i) The pollution of sample is normal

(ii) Since treatment effects are additive, so in a two way classification the observed value is

$\frac{2.527}{1.113} = 2.27$

Where μ is the mean of population, α_i, the effect due to first factor, β_j is effect due to second factor and e_{ij} is the error term due to unknown factors.

(iii) Randomly selected individuals from the population.

(iv)The variance between the samples is constant.

It is supposed that apart from affecting the mean of the samples different treatments do not change the variance of the samples.

Here the calculation of F is based upon the above four assumptions.

15.7 SOLVED PROBLEMS

(1) Three varieties of coal were analysed by four chemists and the ash-content in the varieties was found to be as given below table.

	Chemists			
Varieties	**1**	**2**	**3**	**4**
A	**8**	**5**	**5**	**7**
B	**7**	**6**	**4**	**4**
C	**3**	**6**	**5**	**4**

Find out the analysis of variance.

Solution: To find out the analysis of variance we form the following tables.

Chemists						
Varieties	1	2	3	4	Total	Squares
A	8	5	5	7	25	625
B	7	6	4	4	21	441
C	3	6	5	4	18	324
Total	18	17	14	15	G=64	1390
Squares	324	289	196	225	1034	

Individual Squares

Chemists				
Varieties	1	2	3	4
A	8	5	5	7
B	7	6	4	4
C	3	6	5	4

Total=366

Test Procedure:

Null Hypothesis:

$$H_0 = \mu_1 = \mu_2 = \mu_3 = \mu$$

$$H_0 = \mu_{0.1} = \mu_{0.2} = \mu_{0.3} = \mu_{0.4} = \mu$$

(i)That is, there is no significant difference between varieties(rows)

(i)That is, there is no significant difference between Chemists(rows)

Alternative Hypothesis H_1:

(i) not all μ_i's equal

(ii) not all μ_j's equal

2. Level of Significance

Let α :0.05

Test Statistics:

Correction Factor(C.F)=$\frac{G^2}{N}=\frac{G^2}{n\times k}$

$=\frac{64^2}{3\times 14}=\frac{1024}{3}$=341.33 $\longrightarrow$(1)

Total Sum of squares= $\sum_{i=1}^{k}\sum_{j=1}^{k}x_{ij}^2-C.F$

=366-341.33

=24.67 $\longrightarrow$(2)

Sum of squares between varieties=$\frac{\sum T_i^2}{4}-C.F$

$=\frac{1390}{4}-341.33$

=347.5-341.33=6.17 $\longrightarrow$(3)

Sum of squares between varieties=$\frac{\sum T_j^2}{3}-C.F$

$=\frac{1034}{3}-341.33$

=344.67-341.33=3.34 $\longrightarrow$(4)

Sum of squares due to error=TSS-SSR

=24.67-6.17-3.34=24.67-9.51=15.16 $\longrightarrow$(5)

ANOVA TABLE

Sources of variation	d.f	S.S	MSS	F-ratio
Between varieties	3-1=2	6.17	3.085	$\frac{3.085}{2.527}=1.22$

Between chemists	4-1=3	3.34	1.113	$\frac{2.527}{1.113} = 2.27$
Error	6	15.16	2.527	-
Total	12-1=11			

Table Value:

(a) Table value of F_e for 2.6 with d.f at 5% level of significance is 5.14

(b)Table value of 6.3 with d.f at 5% level of significance is 8.94.

Now Inference is

(i) Since the calculated F_0 is less than table of F_e , We may accept our H_0 for between varieties and say that there is no significant difference between varieties.

(ii) Since calculated F_0 is less than table value of F_e ,for chemists, We may accept our H_0 and may say that there is no significant difference between chemists.

(2) In an experiment on the effects of temperature conditions in human performance, 8 practised Subjects were given a Sensorimotor test in each of 4 temperature Conditions. Since the subjects were all practised, the order in which the teste were done was unimportant. The tests were randomized amongst the subjects, so that for each condition there were equal number of first testing, second testing, third testing and fourth testing. The scores in the tests are shown in the following table.

Temperature(B)	**SUBJECT-(A)**							
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
1	**76**	**80**	**79**	**90**	**85**	**101**	**94**	**83**
2	**75**	**81**	**77**	**90**	**86**	**98**	**93**	**85**
3	**76**	**78**	**66**	**91**	**82**	**98**	**92**	**83**
4	**68**	**75**	**72**	**85**	**82**	**90**	**82**	**77**

Perform an analysis of variance and show whether there is any Significant difference between the Scores of subjects due to temperature Conditions:

Solution: solve this, we can take working mean at x = 84, that is, we get $\mu_{ij} = (x_{ij} - 84)$.

Then the calculations of variances are shown below table:

	SUBJECT-(A)								S_j	S_j^2	μ_{ij}^2
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**			
Temperatures(B)	**1**	-8 (64)	-4 (16)	-5 (25)	6 (36)	1 (1)	17 (289)	10 (100)	-1 (1)	16	256	532
	2	-9 (81)	-3 (9)	-7 (49)	6 (36)	2 (4)	14 (196)	9 (81)	1 (1)	13	169	457
	3	-8(64)	-6 (36)	-8 (64)	7 (49)	-2 (4)	14 (196)	8 (64)	1 (1)	4	16	478
	4	-16 (256)	-9 (81)	12 (144)	1 (1)	-2 (4)	6 (36)	-2 (4)	-7 (49)	-41	1681	575
S_i		-41	-22	-20	-32	-1	51	25	-8	S= -8	2122	2042
S_i^2		1681	484	1024	400	1	2601	625	64	6880		
μ_{ij}^2		465	41	282	122	13	217	249	52	2042	$\sum_i \sum_j \mu_{ij}$	

(i) The total sum of squares= $\sum_i \sum_j \mu_{ij}^2 - \frac{S^2}{N}$

$$=2042 - \frac{64}{32} = 2042-2=2040$$

(ii) Sum of the squares between subjects= $\frac{\sum S_i^2}{n_i} - \frac{S^2}{N} = \frac{6880}{4} - \frac{64}{32} = 1720 - 2 = 1718$

(iii Sum of the squares between temperatures

$$= \sum_j \frac{S_i^2}{n_j} - \frac{S^2}{N} = \frac{2122}{3} - 2 = 263.25$$

(iv) Residual sum of squares=2040-1718-263.25=53.75

The analysis of variance table

Source of variation	Sum of squares	Degrees of freedom	Estimate of variance	F
Between subjects	1718	7	$\frac{1718}{7} = 245.4$	$\frac{245.4}{2.8} = 87.64$

Between Temperatures	263.25	3	$\frac{263.25}{3} = 87.75$	$\frac{87.75}{2.8} = 31.34$
Residual	58.75	21	$\frac{58.75}{21} = 2.8$	-
Total	2040	31	-	-

The 5% value of F for V_1=7 and V_2=21 is 2.50, and for V_1=3 and V_2=21

The value of F=3.07

Since both the calculated values much greater than the corresponding tabular values, we reject the hypothesis that the data are homogeneous.

The Variance between subjects is very large, which shown that there are highly Significant differences in ability between the subjects.

The effects of temperature conditions are also significant.

(3) In one way classification, if the mean sum of Squares between groups is found to be less than that with in groups, would you infer that there are no Significant difference among the groups or would you suspect a further source of variation?

Solution: Let $s_t^2 = \frac{1}{k-1}\sum_{i=1}^{k} n_i(\overline{x_i} - \bar{x})^2$

$$s_p^2 = \frac{1}{n-k}\sum_{i=1}^{k}\sum_{j=1}^{n_i} (x_{ij} - \overline{x_i})^2$$

Where s_t^2 is the mean sum of squares between groups, s_p^2 the mean sum of squares within the groups, then $s_t^2 \leq s_p^2$, that is, F≤ 1, then the dispersion of the sample means about the general mean is not greater than expected ordinarily from a Set of ' k' samples taken from the same populations.

Then it is taken to be insignificant and no further test is necessary.

However $s_t^2 > s_p^2$, that is F > 1; We will apply a variance ratio test of Significance.

15.8 SUMMARY

Analysis of Variance involves the use of variance for testing the significance of the difference between two or more samples under study. Other than 't' and z tests, the test carried out is

called the F test that can be done even when there are more than two groups of Samples from the Population. It is more useful method than t and z tests.

15.9 TECHNICAL TERMS

- One-Way Classification
- Fixed Effect Model
- Two-Way Classification
- F-Statistic
- Degrees of Freedom
- Homogeneity of Variances
- Normality Assumption

15.10 SELF-ASSESSMENT QUESTIONS

SHORT:

1. What is one-way classification in ANOVA, and what distinguishes it from other classification types?
2. Define a fixed effect model in the context of ANOVA.
3. What is two-way classification in ANOVA, and when is it used?
4. List the key assumptions that must be satisfied for a valid ANOVA test.
5. Explain the role and interpretation of the F-statistic in ANOVA.

ESSAY:

(1) What are the purposes of analysis of variance and what are the assumptions involved in the interpretation of such analysis.

Three Variances A, B, C of a crop are tested in a randomized block design with four replications, the layout being given in the diagram appended. The plot yields in pounds are also indicated there in. Analyse the experimental yield, and state your Conclusion (5% Value of F for $V_1 = 2$ and $V_2=9$ is 4.26).

(2) Set up a two way ANOVA table for the data given below

Pieces of field	Treatment			
	ABCD			
P	45	40	38	37
Q	43	41	45	38
R	39	39	4	41

(Row means are equal. Treatment do not differ significantly).

(3) The following table gives the yields from 4 varieties of wheat sown each in 5 plots, all of approximately equal fertility. How will you test the hypothesis that the varieties are not significantly different? Shift the data to a suitable arbitrary origin.

A	32	34	33	35	37
B	34	33	36	37	35
C	31	34	35	32	36
D	29	26	30	28	29

[S.S within varieties 51.20 with 16 df S.S between varieties 134.0 with 3df

$F = \frac{134.0/5}{51.2/16} = 13.9$, Table values for V_1=3, V=16 is 3.24]

4) Give the linear model a for two way classification and derive the analysis of variance by method of least squares.

15.11 FURTHER READING

(1) "Introduction to probability and statistics" by J. Susan Milton and J.C. Arnold, 4th edition, TMH (2007)

(2) "Mathematical Statistics" by R.K. Goyal, Krishna Prakashan Media (P) Ltd, Meerut.

(3) "Fundamentals of Mathematical Statistics" by S.C. Gupta and V.K. Kapoor, S.Chand & Sons, New Delhi

Dr. T.V. Pradeep Kumar

LESSON - 16

RANDOMISED COMPLETE BLOCK DESIGN

OBJECTIVE:

After studying the lesson the students are expected to have clear comprehension of the theory and practical utility of the concepts of measures of Randomised complete Block Design and their area of applications.

STRUCTURE OF THE LESSON:

16.1 Introduction

16.2 Randomised Complete Block Design (RCBD)

16.3 Models for RCBD

16.4 Parameter Estimates for RCBD models

16.5 RCBD model Advantages and Disadvantages

16.6 Applications of RCBD

16.7 Worked out Examples

16.8 Exercise

16.9 Summary

16.10 Technical Terms

16.11 Self-Assessment Questions

16.12 Further Reading

16.1 INTRODUCTION

As already pointed out, statistics deal with methods of collection, tabulation, analysis and interpretation of data. The uncertainty of the results based on the data is explained by the mathematics of probability. It is thus clear that the necessity is the correct method of collection of data. There are some underlying assumptions to every analysis and it is for the investigator to see that these experiment is performed in a manner so that these assumptions are satisfied. The complete sequence of the steps taken to ensure an objective analysis leading to valid references is called “The Design of Experiment” and it is an important step in statistical analysis. How heterogeneity of experimental units can reduce the sensitivity of an experiment, How the Randomised

Complete Block Design (RCBD) can be used to reduce the heterogeneity of the experimental units, how to conduct the analysis of variance (ANOVA) for an experiment that employs the RCBD and how to test for the efficiency of the RCBD versus that of the completely Randomised design. However, RCBD is used to control or handle some systematic and random sources (nuisaver Factor) of variation if they exist.

Randomised Complete Block Design (RCBD) is arguably the most common design of experiments in many disciplines, including agriculture, engineering, medical, etc. In addition to the experimental error reducing ability, the design wide as the generalization of the study findings. For Example, if the study contains the place as a blocking factor, the results could be generalised for the places. A fertilizer producer can only claim that it is effective regardless of the climate conditions when it is tested in various climate conditions. In this lesson we discussed the need for RCBD, its definition. With specific applications.

16.2 RANDOMISED COMPLETE BLOCK DESIGN (RCBD):

The RCBD is also known as the randomised block design in an RCBD experimental units are grouped into blocks and treatments are randomly assigned to the units within each block. The term complete block refers to the facts.

A block is a set of experimental units that are homogeneous in some sense. Hopefully, units in the some block will have similar responses if applied with the same treatment block designs. Which randomise the units within each block to the treatments. In RCBD if want to test treatments. There are b blocks of units available, each block contains k = rg units. Then that each treatment combination is applied to all blocks.

1. Within each block, the k = rg units are randomised to the g treatments r units each.
2. “Complete “means each of the g treatments appears the same number of times(r) in every block.
3. Mostly, block size k=r of treatments g i.e.; r=1.
4. Matched pair design in a special case of RCBD in which the block size k=2.

	Block 1	Block 2	……. ……	Block b
Treatment 1	Y_{11}	Y_{12}	……. ….	Y_{1b}
Treatment 2	Y_{21}	Y_{22}	……. …	Y_{2b}
……. ……	……. ….	……. ….	……. ….	……. …..
Treatment g	Y_{g1}	Y_{g2}	……. …..	Y_{gb}

Normally data are shown arranged by block and treatment. Cannot tell from the data what was not randomised

Advantages of Blocking:

Blocking is the second basic principle of experimental design after randomisation. Blocking what we can,randomise everything else. If units are highly variable, grouping them into more similar blocks can lead to a large increase in efficiency i.e., more power to detect difference in treatment effects. Therefore, this choice of blocks is crucial. Commonly used blocking is (a)Block on batch (for example: Milk produced in a day). (b) Block spatially (c) block on time (d) block when you can identify a source of variation for example, age, gender and history blocks.

16.3 MODELS FOR RCBDS:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk}$$

In which Y_{ij} =response of the unit receiving treatment i in block j

μ= The grand mean

α_i =The treatment effects

β_j =The blocks effects

ε_{ij} =Measurement error, i.e., d~N(0,σ^2)

Just like the models for factorial design,the model above is over parameterized. Some constraints must be imposed. The commonly used constraints are

$$\sum_{i=1}^{3}\alpha_i = 0 \ and \ \sum_{j=1}^{5}\beta_j = 0$$

16.4 PARAMETER ESTIMATES FOR RCBD MODELS:

The model for a RCBD

$Y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$ ε_{ij}'s are I.e; d N(o,σ^2)

It has the same form as the additive model for a balanced two-way factorial design. So the two models have the same least square estimates for parameters.

$$\hat{\mu} = \overline{Y}..$$

$$\hat{\alpha}_i = \overline{Y}_i. - \hat{\mu} = \overline{Y}_i. - \overline{Y}..$$

$$\hat{\beta}_j = \overline{Y}._j - \hat{\mu} = \overline{Y}._j - \overline{Y}..$$

for i = 1, ……..g and j = 1……..b

Fitting values for RCBD:

The fitted values for RCBD is then

$$\hat{Y}_{ij} = \hat{\mu} + \hat{\alpha}_j + \hat{\beta}_j$$

$$= \overline{Y}.. + (\overline{Y}_i. - \overline{Y}..) + (\overline{Y}.j - \overline{Y}..)$$

$$= \overline{Y}_i. + \overline{Y}._j - \overline{Y}..$$

==> row mean + column mean - grand mean

Like an additive model for a balanced two - way design

	Block 1	Block 2	……,……..	Block b	Row mean
Treatment 1	Y_{11}	Y_{12}	……. …	Y_{1b}	$\overline{Y}_{1.}$
Treatment2	Y_{21}	Y_{22}	……. …	Y_{2b}	$\overline{Y}_{2.}$
……. …..	……. ….	……. …	……. …	……. …	……. …
Trestment g	Y_{g1}	Y_{g2}	……. …	Y_{gb}	$\overline{Y}_{g.}$
Column mean	$\overline{Y}_{.1}$	$\overline{Y}_{.2}$	……. …	$\overline{Y}_{.b}$	Grand mean $\overline{Y}_{..}$

Sum of squares and degrees of freedom:

The sum of squares and degrees of freedom for

$$MS_{trt} = \frac{S.S_{trt}}{g-1}, MS_{block} = \frac{SS_{block}}{b-1}, MSE = \frac{SSE}{(g-1)(b-1)}$$

Under the effects model for RCBD.

$$Y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij} \ , \ \varepsilon_{ij}\text{'s are I.e; d } N(o,\sigma^2)$$

Where one can show that

$$E(MS_{trt}) = \sigma^2 + \frac{b}{g-1}\sum_{i=1}^{g}\alpha_i^{\,2}$$

$$E(MS_{block}) = \sigma^2 + \frac{g}{b-1}\sum_{j=1}^{b}\beta_j^{\,2}$$

$$E(MSE) = \sigma^2$$

Thus, MSE is an unbaised estimator of σ^2

ANOVA F- TEST FOR TREATMENT EFFECT:

To test whether there is an treatment effect. $H_0 : \alpha_1 = \alpha_2 = \ldots\ldots\ldots\ldots = \alpha_g$ *vs* H_1 : not all α_i'*s* is are equal.Which is equivalent to testing whether all α_i'*s* are zero.

$H_0 : \alpha_1 = \alpha_2 = \ldots\ldots\ldots\ldots = \alpha_g = 0$ vs H_1: not all α_i'*s* are zero because of the constraint

$$\sum_{i=1}^{g}\alpha_i = 0$$

This test is also equivalent to a comparison between 2models

complete model: $Y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$

Reduced model: $Y_{ij} = \mu + \beta_j + \varepsilon_{ij}$

The test statistic is

$$F_{trt} = \frac{MS_{trt}}{MSE} \sim F_{g-1,(g-1)(b-1)} \text{ under } h_0$$

ANOVA TABLE FOR RCBD:

Source of variation	Degrees of Freedom	Sum of squares	Mean sum of squares	F-test statistic
Treatment	g-1	SS_{trt}	MSS_{trt}	$F_{trt} = MS_{trt}/MSE$
Block	b-1	SS_{block}	MSS_{block}	$F_{block}=MS_{block}/MSE$
Error	(b-1)(g-1)	SSE	MSE	
Total	(bg-1)	SST		

The F statistic F_{block} for testing the block effect, is being omitted since not of much interest.

ANOVA TABLE FO CRD AND RCBD:

As we after ignoring the block effect and analysed the experiment as a CRD, the ANOVA table becomes

SOURCE OF VARIATION	DEGREES OF FREEDOM	SUM OF SQUARES	MEAN SUM OF SQUARES	F- TEST STATSTIC
Treatment	g-1	SS_{trt}	MS_{trt}	$F_{trt}= MS_{trt}/ MSE_{CRD}$
Error	bg-g	SSE_{CRD}	MSE_{CRD}	
TOTAL	bg-1	SST		

In the table SS_{trt} *and* MS_{trt} is the same in CRD and RBCD, but the variability due to block is now in the error term.

$SSE_{CRD} = SSE_{RCBD}+SS_{block.}$

If SS_{block} is large, by considering the block effect, one can substantially reduce the size of noise with a smaller MSE, it is easier to detect difference in treatment effects.

RCBD can be a very effecting noise- reducing technique if SS_{block} is large.

16.5 RCBD MODEL ADVANTAGES AND DISADVANTAGES:

The model RCBD has the following advantages and disadvantages.

Advantages:

1. RCBD's are more proximate than the completely randomised design (CRD) because of grouping.
2. RCBD's can incorporate any numbers of treatments and blocks and replication of some treatments can be done more than others.
3. Even with missing data, the statistical analysis is relatively easier.
4. Estimation of missing plots can be done easily.
5. The calculation of unbiased error can be for specific treatments.

Disadvantages:

1. With large number of treatments RCBD is not suitable I.e, if there are too many treatments, blocks can become too large.
2. The model is not suitable for high variability in complete blocks. I.e, RCBD's are not suitable when there is a lot of variability in complete blocks.
3. Interactions between block and treatment effects can increase error.

Hence, RCBD's are useful for comparing treatment means when there is a single source of variability therefore, they are one of the most wisley used designs and are appropriate when they can control variation in an experiment.

16.6 APPLICATION OF RCBD:

The randomised complete block design(RCBD) is also known as the randomized block design. The randomised complete block design (RCBD) is a common experimental design used in many fields, including agricultural engineering and medicine. It is used to reduce bias and errors and to improve the robustness of statistical analysis. Here are some applications of RCBD.

(a) Agricultural Experiments:

RCBD is the standard design for agricultural experiments, where similar experimental units are grouped into blocks. In a field experiment, treatments are blocked perpendicular to a gradient in the field, such as soil properties or previous crop history.

(b) Pharmaceutical Studies:

RCBD can be used in pharmaceutical studies to reduce bias, errors, and variability within treatment conditions.

(c) Animal Science:

RCBD can be used in animal science to reduce bias, errors, and variability within treatment conditions.

(d) Machine learning and software engineering:

Randomised complete block design (RCBD) is a method of controlling systematic variations in experiments by dividing experimental units into groups or blocks and randomly assigning treatments to units within each block. Hence it is used in many fields and can be applied to machine learning and software engineering by controlling the nuisance factors.

16.7 WORKED OUT EXAMPLES:

Example 1: A study is conducted to compare 4 formulations of a new drug in terms of the availability of the drug in the blood stream overtime.Ten healthy subjects are selected and each subject recieves each drug in random order in a randomised block design. The researchers conducts the appropriate F-Test for testing for formulation differences. If the test is conducted of the alpha = o.o5 significance level he will conclude formulation differences exist if the F- Statistic falls in what range?

Sol: $df_{tr} = 4\text{-}1 = 3,\ df_{Err} = (10\text{-}1)(4\text{-}1) = 27\quad f_{0.05,3,27} = 2.960$

Example 2:

Compute an ANOVA table using (Randomised complete block design) RCBD for six varieties of maize in three blocks.

Variety	1	2	3
1	2.5	2.7	2.9
2	3.8	3.9	2.9
3	5.8	5.9	5.7
4	1.2	1.4	1.3
5	4.5	4.8	4.9
6	1.0	1.1	1.2

Sol: We were the above table

Variety	1	2	3	Row Total
v_1	2.5	2.7	2.9	8.1
v_2	3.8	3.9	2.9	11.9
v_3	5.8	5.9	5.7	17.4
v_4	1.2	1.4	1.3	3.9
v_5	4.5	4.8	4.9	14.2
v_6	1.0	1.1	1.2	3.3
Column Total	18.8	19.8	20.2	58.8

Correlation factor(C.F) $= \frac{\sum x}{n} = \frac{(58.8)^2}{18} = \frac{3457.44}{18} = 192.08$

Total sum of squares = Sum of squares of each observation - C.F

= (

$2.5^2 + 2.7^2 + 2.9^2 + 3.8^2 + 3.9^2 + 4.2^2 + 5.8^2 + 5.9^2 + 5.7^2 + 1.2^2 + 1.4^2 + 1.3^2 + 4.5^2 + 4.8^2 + 4.9^2 + 1.0^2 + 1.1^2 + 1.3^2$
) - C.F

TSS = 246.22 - 192.08 = 54.14

Variety sum of squares VSS = $\frac{\sum (Y_i)^2}{n} - C.F$

$$\frac{\left[8.1^2 + 11.9^2 + 17.4^2 + 3.9^2 + 14.2^2 + 3.3^2\right]}{3} - 192.08$$

VSS = [737.72/3] - 192.08

VSS = 245.9 - 192.08

VSS = 53.83

Block sum of squares (BSS) $= \left[\frac{18.8^2 + 19.8^2 + 20.2^2}{6}\right] - 192.08$

I.e; BSS = $\frac{\sum x^2}{n} - C.F = 192.25 - 192.08 = 0.17$

Error sum of square (ESS) = TSS - (VSS + BSS)

==> 54.14 - 583 + 0.117

ESS = 0.14

ANOVA Table

Source of variation : df		S.S	M.S	Fcal	$F-tab$ 5%	1%
Total	17	54.14	3.18			
Variety	5	53.83	11.77	840.71	3.33	5.64
Block	2	0.17	0.09	6.43	4.10	7.68
Error	10	0.14	0.014			

I.e; (d.f) Degrees of freedom = n - 1

T.D.F= 18 - 1 = 17

variety d.f = 6 - 1 = 5

Block d.f = 3 - 1 = 2

Error df = Pdf - (Vdf + Bdf) = 17 - (5 +2) = 10

Mean sum of squares due to total TMS = TSS/Tdf = 54.14/17 = 3.18

Mean sum of squares due to varieties VMS = VSS/Vdf = 53.83/5 = 11.77

Mean sum of squares due to blocks = BSS/Bdf = 0.17/2 = 0.09

Mean sum of squares due to errors = ESS/Edf = 0.14/10 = 0.014

F Tabulated values due to variety at 5% is 3.33

1% is 5.64

F Tabulated values due to block at 5% is 4.10

1% is 7.68

Taken from F- statistical tables

F calculated value due to varieties = variety Mean square/Error Mean square

==> 11.77/0.014 = 840.71

F calculated value due to block = Block Mean square/Error Mean square

==> 0.09/0.014 = 6.43

Example 3:

consider the results given in the following table for an experiment involving six treatments in four randomised blocks. The treatments are indicated by numbers

Yield for a randomised block experiment

Block	Treatment yield					
1	24.7	27.7	20.6	16.2	16.2	24.9
2	22.7	28.8	27.3	150	22.5	17.0

3	26.3	19.6	38.5	36.8	39.5	15.4
4	17.7	310	28.5	14.1	34.9	22.6

Test whether the treatments differ significantly. Also

(i) determine the critical difference between the means of any two treatments, and obtain the efficiency of this design relative to its layout.

Solution: Null Hypothesis: $H_t : t_1 = t_2 = \ldots\ldots\ldots = t_6$

$H_b : b_1 = b_2 = b_3 = b_4$

i.e; the treatments as well as blocks are homogeneous for finding various sum of B_j squares we rearrange the above table as follows.

Blocks	Treatments						block totals(B_j)	B_j
1	24.7	20.6	27.7	16.2	16.2	24.9	130.0	36900
2	27.3	28.8	22.9	15.0	17.0	22.5	133.3.	17768.89
3	38.5	39.5	36.8	19.6	15.4	26.3	176.1	31011.27
4	28.5	31.0	34.5	14.1	17.7	22.6	148.8.	22.14.44
Treatment Totals								
(T2)	119.0	119.9	122.1	64.9	66.3	96.3	388.5 = G	
T_d^2	14161	14376.01	14908.41	4212.01	4395.69	9273.69		
Average	29.75	30.0	30.5	16.2	16.6	24.1		

Correction Factor = $\frac{G^2}{n}$ = 346332.25/24 = 14430.57

Raw sum of squares = $\sum\sum y_{ij}^{\ 2}$ = 15780.76

Total sum of squares = RSS-C.F = 15780.76 - 14430.57

==> 1350.25

SS due to treatments (SST) = $\frac{1}{4}\sum_{i=1}^{6} T_i^2 - C.F$

= $\frac{61326.81}{4}$ - 14430.57 = 901.19

Sum of squares due to blocks (SSB) = $\frac{1}{6}\sum_{j=1}^{4} B_j - C.F$ = $\frac{87899.63}{6}$ - 14430.57 = 219.63

Error sum of squares = TSS - SST - SSB

= 1350.25 - 901.19 - 219.43 = 229.63

Analysis of variance table (ANOVA):

Source of variation	d.f	sum of squares	Mean sum of squares(MSS)	Variance ratio(F)
Treatment	5	901.19	180.24	Ft=180.24/15.31=11.8
Block	3	219.43	73.14	Fb=73.14/15.31 = 4.7
Error	15	229.63	15.31	
Total	23	1350.25		

F-tabulated values:

$F_{0.05}(3,15) = 5.42$ *and* $F_{0.05}(5,15) = 4.5$

Conclusion: Since , $F_{tab.val}$ due to treatments = 11.8 > $F_{tab.val}$ = 5.42 at 5% l.o.s we reject the null hypothesis H_t and conclude that the treatments differ significantly. Again, Fcal value due to blocks = 4.7 > $F_{tab.val}$ and we conclude that the blocks differ significantly.

As $H_t : t_1 = t_2 = ... = t_6$ is rejected, we are interested to find which treatment means differ significantly.

Critical difference for any two treatment means

= $t_{0.05}$ *for error d.f* $X\sqrt{\frac{2s_E^{\,2}}{r}}$ Where 'r' is the number of times a treatment is replicated.

= 2.131 x $\sqrt{\frac{2\times 15.31}{4}}$ = 2.131 x 2.8 = 5.97 (Approx)

By comparing the differences between the mean yields for different treatments with the critical difference, we find that treatments 3,2, and 1 are alike in giving significantly high yields while treatments 4 and 5 are alike in giving significantly low yields. By using formula, the efficiency of the above RCBD relative to its layout as CRD is given by

$$E = \frac{r(t-1)s_E^{\,2} + (r-1)S_B^{\,2}}{(rt-1)\,s_E^{\,2}}$$

Where r is the number of replicates (blocks) and t is the number of treatments. After substituting the values we get

$$E = \frac{4\times5\times15.31+3\times73.14}{23\times15.31} = 1.49$$

Hence gain in efficiency is 49%

Example 4:

A randomised complete block design is conducted to compare the output of three weaving looms (treatments) for a sample of 10 operators (blocks) , where each operators output is measured on each loom. The mean square error from the ANOVA is MSE = 500. Compute Bonferrons B, the minimum significant difference for concluding that two looms population means differ if their sample means differ by at least B.

Solution: Here treatments $t_r = 3$ b = 10 MSE = 500

d.f due to error = (3-1)(10-1) = 18, $c = \frac{3(3-1)}{2} = 3$

comparison.

$t_{0.025}$, c=3 , d.f = 2.639 , $B = 2.639\sqrt{\frac{2(500)}{10}} = 26.39$

Example 5:

A college's volleyball coach is interested in whether her team player's prefer one brand of shoe among 3 brands. She has each player practice wearing each brand (in random order), and has each player rate the comfort on a 0-100 visual analogue scale. She analyses the experiment as a Randomised complete block. Design with t = 3 treatments (shoe brands) and b = 18 (team players). She obtains the following results from the analysis of variance

$\overline{Y}_1 = 65$, $\overline{Y}_2 = 80$, $\overline{Y}_3 = 55$, MSE=225

a. Use Bonferron's method to obtain simultaneous 95% confidence intervals for the population, mean rating differences among all pairs of brands

$df_{error} = 2(17) = 34$ c = 3 $t_{0.025,3,34} = 2.518$ $B = 2.518\sqrt{\frac{2(225)}{18}}$

==> 12.59

$(\mu_1 - \mu_2)$: (65 - 85) ± 12.6 = (-27.6, -2.4), $(\mu_1 - \mu_3)$: (-2.6,22.6)

$(\mu_2 - \mu_3)$: (12.4,37.6)

b. Show the results by drawing lines that join means that are not significantly different: 3 1 2

Example 6: A randomised complete Block Design was conducted to compare 3 energy drinks in terms of endurance in 6 subjects (each subject drinks each energy drink). The response is time to exhaustion on a treadmill. Due to some outlying observations, use Friedman's test to determine whether the 3 energy drinks differ in their effects on endurance.

Subject	Drink1	Drink2	Drink3
1	42(1)	48(2)	62(3)
2	36(2)	34(1)	48(3)
3	54(1)	56(2)	75(3)
4	44(1)	46(2)	52(3)
5	28(1)	32(2)	44(3)
6	45(1)	50(2)	65(3)

T_1 = 1+2+1+1+1+1 = 7 T_2 = 2+1+2+2+2+2 = 11, T_3 = 6(3) = 18

Solution

a. The test statistic: $F_r = \frac{12}{6(3)(3+1)}\left[7^2 + 11^2 + 18^2\right]$ - 3(6)(3+1)

$$==> \frac{12(494)}{72} - 72$$

==> 82.333 - 72

==> 10.333

b. Rejection region is:

$$F_r \geq \chi^2_{0.05,3-1} = 5.991$$

Example 7:

Grain yield of rice at six seeding rates(Mg/ha)

	Seeding rate (kg/ha)						
Rep	25	50	75	100	125	150	y.j

1	5.1	5.3	5.3	5.2	4.8	5.3.	31.0
2	5.4	6.0	5.7	4.8	4.8	4.5.	31.2
3	5.3	4.7	5.5	5.0	4.4	4.9.	29.8
4	4.7	4.3	4.7	4.4	4.7	4.1.	26.9
Y_i	20.5	20.3	21.2	19.4	18.7	18.8.	118.9
$\sum Y_{ij}^2$	105.35	104.67	112.92	94.44	87.53	89.16	594.05

Solution:

Step 1: Calculate the correction factor (CF)

$$\text{CF} = \frac{Y..^2}{tr} = \frac{118.9^2}{6.4} = 589.050$$

Step 2: Calculate the total sum of squares (TSS)

$$\text{Total SS} \quad \sum Y_{ij}^2 - CF$$
$$= \left[5.1^2 + 5.4^2 + 5.3^2 + ... + 4.1^2\right] - \text{C.F}$$

$$= 5.02$$

Step 3: Calculate the replicate SS

$$\text{Rep SS} = \sum \frac{Y_j^2}{t} - C.F$$

$$= \frac{\left[31.0^2 + 31.2^2 + 29.8^2 + 26.9^2\right]}{6} - CF$$

$$= 1.965$$

Step 4: Calculate the treatment SS ($T_{rt}\ SS$)

$$T_{rt}\ SS = \sum \frac{Y_i.^2}{r} - CF$$

$$= [\,20.5^2 + 20.3^2 + 21.2^2 + 19.4^2 + 18.7^2 + 18.8^2\,]/4 \text{ - CF}$$

$$\Longrightarrow 1.2675$$

Step 5: Calculate the error SS

Error SS = Total SS - Rep SS - Trt SS

$$\Longrightarrow 5.02 - 1.965 - 1.2675$$

$$\Longrightarrow 1.7875$$

Step 6: Construct the ANOVA table

Source of Variation	DF	SS	MS	F-ratio
Replication	r-1 = 3	1.9650	0.6550	RepMS/ErrorMS = 5.495
Treatments	t-1 = 5	1.2675	0.2535	Trt MS/Error MS = 2.127
Error	(r-1)(t-1) = 15	1.7875	0.1192	
Total	tr-1 = 23	5.0200		

Step 7: F-table values for Replications and treatments

For Replications: $F_{0.05;3,15} = 3.29$

$F_{0.01;3,15} = 5.42$

For treatments: $F_{0.05;5,15} = 2.90$

$F_{0.01;5,15} = 4.56$

Step 8: We make conclusions as in case of replication

Since $F_{cal\ value}$ 5.495 > $F_{\text{table } value}$ at 5% and 1% level of significance , we reject H_o: all replicate means are equal.

For treatments: Since $F_{cal\ value}$ 2.127 < $F_{\text{table } value}$ the at 5% and 1% level of significance we accept null hypothesis H_o: All treatments means are equal.

Step 9: Coefficient of variation is

CV = $\frac{SD}{\bar{Y}} \times 100$ SD: Standard Deviation

CV = $\sqrt{\frac{0.1192}{4.95}}$ x 100 $\bar{Y}$: Sample mean

CV = 6.97%

16.8EXERCISE:

The following table gives the gain in weights of four different types of pigs fed on three different rations. Test to see whether the rations or the pigs types differ in their effects on mean weights.

Types of pigs		I	II	III	IV
	I	7	16	10	11
Types of	II	14	15	15	14
Rations	III	8	16	7	11

Ans: 4.94 , 9.2

2. Randomised complete design with the five treatments O,A,B,C,D and 4 blocks was used. The plan and yields in lbs.per plot were as follows.

Block I	D 67	B 69	A 70	C 64	O 65
Block II	B 71	C 69	A 73	O 69	D 68
Block III	O 71	D 70	C 69	A 75	B 71
Block IV	C 67	A 70	O 63	B 69	D 71

Prepare the analysis of variance table and test the homogeneity of means between treatments and blocks. Also give standard error between any two treatments means and between any two block means. Also find out efficiency of this arrangements

3. The following is the layout and yields in kgs of 4 varieties of what in 4 blocks.

	I	A5	C13	D7	B11
Block	II	B12	A6	D8	C13
	III	D7	C15	A6	B12
	IV	C14	A8	B13	D9

Perform an analysis of this data and interpret the results.

4. Three varieties A, B, C of a crop are tested with four replications each. The layout and the plot yields in pounds are given below. Analysis the experimental yield and state your conclusions.

5.

I	II	III	IV
A - 6	C - 5	A - 8	B - 9
C - 8	A - 4	B - 6	C - 9
B - 7	B - 7	C - 10	A - 6

Also examine:

(i) If I and III replicates (Blocks) are significantly different ,

(ii) which of the varieties gives the maximum mean yields?

6. Discuss the advantages and disadvantages of RCBD.

7. Analyse the following data and comment on your findings.

	Treatments			
Blocks	A	B	C	D
1	300	330	360	290
2	240	230	350	240
3	370	360	390	310
4	270	320	320	260

8. Give the layout of RCBD and explain the Situations where it is used.

9. Analyse the following data. Also apply t-test to examine if the (i) 2nd and 4th treatments and (ii) 1st and 3rd block effects are significantly different.

	T_1	T_2	T_3	T_4
B1	18	30	24	21
B2	28	32	29	30
B3	17	28	21	26

10. Explain the importance of RCBD.

11. Analyse the following data and give your comments.

Group I	560	600	500	650	640
Group II	480	610	480	520	460
Group III	550	600	440	460	550

16.9 SUMMARY:

In this lesson an attempt is made to explain the concept of randomised complete block design - layout, analysis, Applications , advantages and disadvantages with both theory and practical. A few examples are worked out and a good number of exercises are also given.

16.10 TECHNICAL TERMS:

- Randomised complete block design (RCBD)
- Treatments
- Complete block
- ANOVA Table.

16.11 SELF-ASSESSMENT QUESTIONS

SHORT:

1. **What is a Randomised Complete Block Design (RCBD)?**
2. **How are treatments defined in the context of an RCBD?**
3. **What constitutes a complete block in RCBD?**
4. **What is the primary purpose of using an RCBD in experiments?**
5. **What information is typically presented in an ANOVA table for an RCBD?**

ESSAY:

1. **Explain the concept of Randomised Complete Block Design (RCBD) and discuss its advantages compared to a completely randomized design.**
2. **Describe the role and importance of treatments in RCBD. How are treatments assigned, and what impact do they have on the experiment's outcome?**
3. **Discuss the concept of a complete block in an RCBD. What criteria define a complete block, and how does it contribute to reducing experimental error?**
4. **Elaborate on the structure of an ANOVA table used in the analysis of an RCBD. Identify the main sources of variation and explain how the F-statistic is used to determine treatment significance.**
5. **Provide a comprehensive example of an RCBD experiment. Describe the experimental design, how blocks and treatments are organized, and demonstrate how to analyze the data using an ANOVA table.**

16.12 FURTHER READING

(1) "Introduction to probability and statistics" by J. Susan Milton and J.C. Arnold, 4th edition, TMH (2007)

(2) "Mathematical Statistics" by R.K. Goyal, Krishna Prakashan Media (P) Ltd, Meerut.

(3) "Fundamentals of Mathematical Statistics" by S.C. Gupta and V.K. Kapoor, S.Chand & Sons, New Delhi

Dr. B. Sri Ram

LESSON- 17

SIMPLE LINEAR REGRESSION

AIMS AND OBJECTIVES:

This lesson is prepared in such a way that a – studying the material the student is expected to have a thorough comprehension of the concepts like simple linear regression which are the important areas of investigation and statistical data analysis. The student will be having and well equipped with both theoretical as well as practical aspects of simple linear regression.

STRUCTURE OF THE LESSON:

17.1 Introduction
17.2 Simple Linear Regression
17.3 Lines of Regression Y on X and X on Y
17.4 Regression coefficients
17.5 Properties of Regression coefficients
17.6 Angle between Two lines of Regression
17.7 Standard Error of Estimate of Residual variance
17.8 Correlation coefficient between observed and Estimated value
17.9 Regression curves
17.10 Worked out Example
17.11 Exercise
17.12 Summary
17.13 Technical Terms.
17.14 Self-Assessment Questions
17.5 Further Reading

17.1 INTRODUCTION

Very often, the interest lies in establishing the octal relationship between two or more variables. This problem is dealt with regression. On the other hand, we are often not interested to know the octal relationship but are only interested in knowing the degree of relationship between two or more variable. This problem is dealt with correction analysis. For both the studies, the number of variables may be two or more.

The concept of regression was first given by Sir Francis Galton (1822-1911) in a study of inheritance of stature in the human being. To prove this biometrical fact, Karl Pearson found the regression of son's height on father's height. But soon the use of regression techniques became too common for a variety of problems. The relationship between variables, if it exists, may be linear or curvilinear.

But scientific, social and economic phenomena do not confine to two variables only. Many studies involve only more than two variables. In these studies, we often need to give actual relationship between three or more variables and 1 or to explain the strength of association between them. For this we want to establish the relationship between the dependent variable and independent variables and a mathematical equation can be given to do this. This type of mathematical equation is known as a mathematical model. The equation pertaining to such as relationship may be of any type. But we deal with a linear relationship which represents a plane according to the number of variables involved.

17.2 SIMPLE LINEAR REGRESSION:

The term "Regression" Literally means "Stepping back towards the average. If there exists some relation between two variables, their scatter diagram should have points clustering near about some curve. If this curve is a straight line, it suggests some linear relationship between the variables and this straight line is known as the line of regression and there is said to be linear regression between the variables, otherwise regression is said to be curvilinear.

The line of regression is the line which gives the best estimate to the value of one variable for any specific value of the other variables. Thus, the line of regression is the line of "best fit" and is obtained by the principle of least squares.

17.3 LINEAR OF REGRESSION Y ON X AND X ON Y:

Let us suppose that in the bivariable distribution $(X_i, Y_i), i = 1, 2, ..., n$, Y is dependent variable and X is independent variable. Let the line of regression of Y on X be $Y = a + bX$. According to the principle of least squares the normal equation for estimating a and b are

$$\sum_{i=1}^{n} y_i = na + b\sum_{i=1}^{n} x_i \quad (1)$$

$$\sum_{i=1}^{n} y_i = na + b\sum_{i=1}^{n} x_i \quad (2)$$

on dividing eq (1) by $\mathbf{n}$,we get

$$\bar{y} = a + b\bar{x}$$

(3)

Thus, the line of regression of y on x passes through the point $(\bar{x}, \bar{y})$. Now

$$\mu_{11} = \text{cov}(x,y) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y} \Rightarrow \frac{1}{n}\sum_{i=1}^{n} x_i y_i = \mu_{11} + \bar{x}\bar{y}$$

(4)

also

$$\sigma_{x^2} = \frac{1}{n}\sum_{i=1}^{n} x_{i^2} - \bar{x}^{-2} \Rightarrow \frac{1}{n}\sum_{i=1}^{n} x_{i^2} = \sigma_{x^2} + \bar{x}^{-2}$$

(5)

Dividing eq (2) by n and using eq(4) and eq(5) , we get

$$\mu_{11} + \bar{x}\bar{y} = a\bar{x} + b(\sigma_{x^2} + \bar{x}^{-2})$$

(6)

Multiplying eq(3) by $\bar{x}$ and then subtracting from eq(6) we get,

$$\mu_{11} = b\sigma_{x^2} \Rightarrow b = \frac{\mu_{11}}{\sigma_{x^2}}$$

(7)

Since b is the slope of the line of regression of Y on X and since the line of regression passes through the point $(\bar{x}, \bar{y})$, its equation is

$$(y - \bar{y}) = b(x - \bar{x}) = \frac{\mu_{11}}{\sigma_{x^2}}(x - \bar{x}) \qquad (\because \text{from eq(7)})$$

(8)

$$\Rightarrow (y - \bar{y}) = r\frac{\sigma_y}{\sigma_x}(x - \bar{x}) \left(\because\ r(x,y) = \frac{\text{cov}(x,y)}{\sigma_x \sigma_y} \Rightarrow \mu_{11} = \text{cov}(x,y) = r\sigma_x\sigma_y \right)$$

(9)

By interchanging the variable X and Y in eq(8) and (9) , the equation of the line of regression of X and Y becomes

$$(x-\bar{x})=\frac{\mu_{11}}{\sigma_{y^2}}(y-\bar{y})$$

(10)

$$\Rightarrow (x-\bar{x})=r\frac{\sigma_x}{\sigma_y}(y-\bar{y})\quad\left(\because \mu_{11}=\frac{\sigma_x}{\sigma_y}r\right)$$

(11)

Note: We always have two lines of regression except case of perfect correlation when both lines coincide, we get only one line. That is, in case of perfect correlation, (r=+1), both the lines of regression coincide.

17.4 REGRESSION COEFFICIENTS:

'**b**' the slope of the line of regression of **Y** on **X** is also called the coefficient of regression of **Y** on **X**. It represents the increment in the value of dependent variable **Y** corresponding to a unit change in the value of independent variable **X**. Then we write

$$b_{yx}=\text{Regression Coefficient of y on x}=\frac{\mu_{11}}{\sigma_{x^2}}=r\frac{\sigma_y}{\sigma_x}$$

(1)

Similarly, the coefficient of regression of **X** on **Y** indicates change in the value of variable **X** corresponding to a unit change in the values of variable **Y** and is given by:

$$b_{xy}=\text{Regression Coefficient of x on y}=\frac{\mu_{11}}{\sigma_{y^2}}=r\frac{\sigma_x}{\sigma_y}$$

(2)

17.5 PROPERTIES OF REGRESSION COEFFICIENTS:

1. Correlation Coefficient is the geometric mean between the regression coefficients i.e.,

$r=\pm\sqrt{b_{xy}b_{yx}}$ we know that

Regression Coefficients of **Y** on **X**

$$r\frac{\sigma_y}{\sigma_x}=b_{yx} \qquad (1)$$

Regression Coefficients of **X** on **Y**

$$r\frac{\sigma_x}{\sigma_y} = b_{xy} \qquad (2)$$

From (1) and (2) we get,

$$r\frac{\sigma_y}{\sigma_x} \cdot r\frac{\sigma_x}{\sigma_y} = b_{yx} \cdot b_{xy} \Rightarrow r^2 = b_{yx} b_{xy} \quad (\because r = \pm\sqrt{b_{yx} b_{xy}})$$

i.e., correlation coefficient $r = \pm\sqrt{b_{yx} b_{xy}}$

2. Regression Coefficients are independent of the change of origin but not of scale i.e., $b_{yx} = \frac{k}{h} b_{vu}$

Let $u = \frac{x-a}{h}, v = \frac{y-b}{k}$

$\Rightarrow x = a + hu \Rightarrow y = b + kv$

Where $a, b, h(>0), k(>0)$ are constants then

$\text{cov}(x, y) = hk\,\text{cov}(u, v), \sigma_{x^2} = h^2\sigma_{u^2}$ and $\sigma_{y^2} = k^2\sigma_{v^2}$

$$b_{yx} = \frac{\mu_{11}}{\sigma_{x^2}} = \frac{hk\,\text{cov}(u, v)}{h^2\sigma_{u^2}} = \frac{k}{h}\frac{\text{cov}(u, v)}{\sigma_{u^2}} = \frac{k}{h} b_{vu}$$

Similarly we can prove that

$$b_{xy} = \frac{h}{k} b_{uv}$$

3. Arithmetic of mean of the regression coefficients is greater than the correlation coefficient r, provided $r > 0$,

i.e., $\frac{1}{2}\left(b_{yx} + b_{xy}\right) \geq r$.

we have to prove that, $\frac{1}{2}\left(b_{yx} + b_{xy}\right) \geq r$

$$\Rightarrow \frac{1}{2}\left(r\frac{\sigma_y}{\sigma_x} + r\frac{\sigma_x}{\sigma_y} \right) \geq r$$

$$\text{or} \quad \frac{\sigma_y}{\sigma_x} + \frac{\sigma_x}{\sigma_y} \geq 2 \ \left(\because r \geq 0 \right)$$

$$\Rightarrow \sigma_{y^2} + \sigma_{x^2} - 2\sigma_x\sigma_y \geq 0$$

$$\text{i.e.,} \left(\sigma_y - \sigma_x \right)^2 \geq 0$$

Which is always true, since the square of a real quantity is ≥ 0

4. If one of the regression coefficients is greater than unity, the other must be less than unity

$$\text{i.e.,} \quad b_{xy} \leq \frac{1}{b_{yx}} < 1$$

Let one of the regression coefficient say b_{yx} be the greater than unity, then we have to show that $b_{xy} < 1$ now $b_{yx} > 1$

$$\Rightarrow \frac{1}{b_{yx}} < 1$$

Also $r^2 \leq 1 \Rightarrow b_{yx} \cdot b_{xy} \leq 1$

$$\text{Hence} \quad b_{xy} \leq \frac{1}{b_{yx}} < 1$$

17.6 ANGLE BETWEEN TWO LINES OF REGRESSION

Equations of the lines of regression of **Y** on **X** and **X** on **Y** are.

$$(y - \bar{y}) = r\frac{\sigma_y}{\sigma_x}(x - \bar{x})$$

and

$$(x - \bar{x}) = r\frac{\sigma_x}{\sigma_y}(y - \bar{y})$$

Slope of these lines are $r\frac{\sigma_y}{\sigma_x}$ and $\frac{\sigma_y}{r\cdot\sigma_x}$ respectively. If θ is the angle between the two lines of regression, then

$$\text{Tan}\,\theta = \frac{r\frac{\sigma_y}{\sigma_x} - \frac{\sigma_y}{r\sigma_x}}{1 + r\frac{\sigma_y}{\sigma_x}\frac{\sigma_y}{r\sigma_x}} = \frac{r^2-1}{r}\left[\frac{\sigma_x\sigma_y}{\sigma_{x^2}+\sigma_{y^2}}\right]$$

$$= \frac{1-r^2}{r}\left[\frac{\sigma_x\sigma_y}{\sigma_{x^2}+\sigma_{y^2}}\right] \qquad (\because r^2 \le 1)$$

$$\theta = \text{Tan}^{-1}\left\{\frac{1-r^2}{r}\left(\frac{\sigma_x\sigma_y}{\sigma_{x^2}+\sigma_{y^2}}\right)\right\}$$

Case (i) : if $r = 0, \tan\theta = \alpha \Rightarrow \theta = \Pi/2$

Thus, if the two variables are uncorrelated, the lines of regression become perpendicular to each other.

Case (ii): if $r = \pm 1, \tan\theta = 0 \Rightarrow \theta = 0 \text{ or } \Pi$

In this case the two lines of regression either coincide or they are parallel to each other. But since both the lines of regression pass through the point $\left(\bar{x}, \bar{y}\right)$ they cannot be parallel. Hence, in the case of perfect conditions, positive or negative, the two lines of regression coincide.

Note:

1. Whenever two lines intersect, there are two angles between them, one acute angle and the other obtuse angle. Further $\tan\theta > 0$ is $0 \le \theta \le \frac{\Pi}{2}$ i.e., θ is the acute angle and $\tan\theta < 0$. If $\frac{\Pi}{2} \le \theta \le \Pi$, i.e., θ is an obtuse angle and since $0 < r^2 < 1$, the acute angle (θ_1) and obtuse angle (θ_2) between the two lines of regression are given by

$$\theta_1 = \text{Actual angle} = \tan^{-1}\left\{\frac{\sigma_x\sigma_y}{\sigma_{x^2}+\sigma_{y^2}}\cdot\frac{1-r^2}{r}\right\}, r>0$$

$$\theta_2 = \text{obtuse angle} = \tan^{-1}\left\{\frac{\sigma_x\cdot\sigma_y}{\sigma_{x^2}+\sigma_{y^2}}\cdot\frac{r^2-1}{r}\right\}, r>0$$

2. When $r=0$ i.e., variable X and Y are uncorrelated, then the lines of regressions of Y on X and X on Y are given respectively by equations (9) and (11) $y=\bar{Y}$ and $x=\bar{X}$ which were shown in the diagram

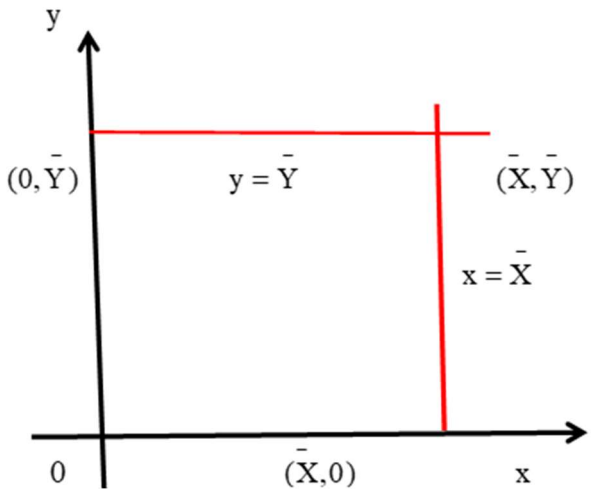
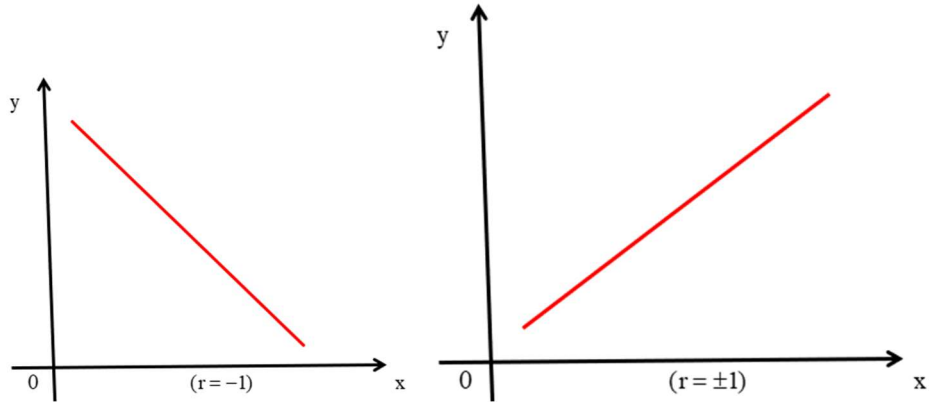


Fig 17.1 $r=0$, the lines of regression are perpendicular to each other

Hence, in this case $r=0$, the lines of regression are perpendicular to each other and are parallel to x-axis and y-axis respectively.

3. The fact that if $r=0$ variable is uncorrelated, the two lines of regression are perpendicular to each other and if $r\pm1$, $\theta=0$ i.e., the two lines coincide, which makes us to conclude that for higher degree of correlation between the variables, the angle between the lines is smaller, i.e., the two lines of regression are nearer to each other. On the other hand, if the lines of regression make a larger angle, they indicate a poor degree of correlation between variables and ultimately for $\theta=\frac{\Pi}{2}$, $r=0$, i.e., the lines become perpendicular if no correlation exists between the variables. Thus, by plotting the lines of regression on a graph paper, we can have an approximate idea about the degree of correlation between the two variables under study. Consider the following illustrations:



1. (b)

Fig 17.2 (a) Two lines coincide $r = -1$ (b) Two lines coincide $r = +1$

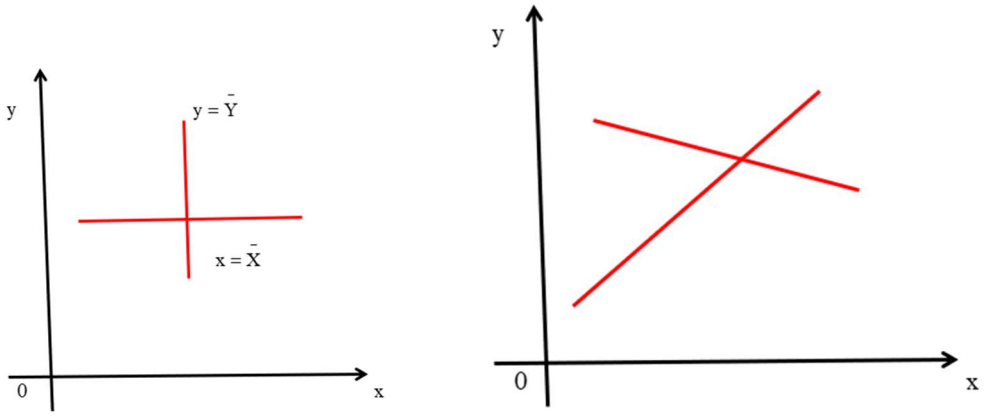


Fig 17.3 (a) Two lines perpendicular $r = 0$ (b) Two lines Apart (i.e., low degree of correlation)

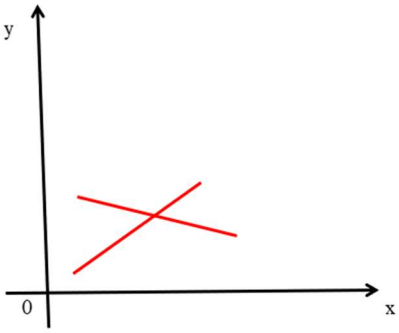


Fig 17.4 Two lines Apart (i.e., High degree of correlation)

17.7 Standard Error of Estimate or Residual variance:

The equation of the line of regression of Y on X is

$$y = \bar{y} + r\frac{\sigma_x}{\sigma_y}(x - \bar{x})$$

$$\Rightarrow \frac{y - \bar{y}}{\sigma_y} = r \cdot \frac{(x - \bar{x})}{\sigma_x}$$

The residual variable S_{y^2} is the expected value of the squares of deviations of the observed values of Y from the expected values as given by the line of regression of Y on X.

Thus,

$$S_{y2} = E\left[y - \left(\bar{y} + \left(r\sigma_y (x - \bar{x}) / \sigma_x \right) \right) \right]^2$$

$$= \sigma_{y^2} \cdot E\left[\frac{(y - \bar{y})}{\sigma_y} - r\left(\frac{x - \bar{x}}{\sigma_x} \right) \right]^2$$

$$= \sigma_{y^2} \cdot E\left[y^* - rx^* \right]^2$$

Where x^* and y^* are standardized variables so that

$$E\left(x^{*2}\right) = 1 = E\left(y^2\right) \qquad \text{and} \qquad E\left(x^* y^*\right) = r$$

$$\therefore S_{y2} = \sigma_{y^2} E\left[E\left(y^{*2}\right) + r^2 E\left(x^{*2}\right) - 2rE\left(x^{*2}\right) \right]$$

$$= \sigma_{y^2}\left(1 - r^2\right)$$

$$\Rightarrow S_{y2} = \sigma_y \left(1 - r^2\right)^{1/2}$$

Similarly, the standard error of estimate of X is given by

$$S_{x2} = \sigma_x \left(1 - r^2\right)^{1/2}$$

Note:

1. Since S_{x2} or $S_{y2} \geq 0$, it follows that $\left(1 - r^2\right) \geq 0 \Rightarrow |r| \leq 1$ hence $-1 \leq r(x, y) \leq 1$
2. If $r \pm 1$, $s_x = 0$ and $s_y = 0$ so that each deviation is zero, and the two lines of regression are coincident.
3. Since, as $r^2 \to 1, s_x \to 0$ and $s_y \to 0$ it follows that departure of the value r^2 from unity indicates the departure of the relationship between the variables X and Y from linearity.
4. From the definition of linear regression, the minima condition implies that s_x and s_y

17.8 Correlation Coefficient between observe and Estimated Value:

Here we will find the correlation between

$$\hat{y} = \hat{y} + r\frac{\sigma_y}{\sigma_x}(x - \bar{x})$$

Where $\hat{y}$ is the estimated value of y as given by the line of regression of y on z , which is given by

$$r(y, \hat{y}) = \frac{r(\hat{y}, y)}{\sigma_y \hat{\sigma}_y}$$

We have

$$E(\hat{y}) = E[\bar{y} + r\frac{\sigma_y}{\sigma_x}(x - \bar{x})]$$

$$= \bar{y} + r\frac{\sigma_y}{\sigma_x}E(x - \bar{x}) = \bar{y}$$

$$\therefore \text{var}(\hat{y}) = E[\hat{y} - E(\hat{y})]^2 = E[r\frac{\sigma_y}{\sigma_x}(x - \bar{x})]^2 = r^2\sigma_{y^2}$$

$$\Rightarrow \hat{\sigma}_y = r\sigma_y$$

$$\text{cov}(y,\hat{y}) = E[y - E(y)][\hat{y} - E(\hat{y})]$$

Also,
$$= E[(b(x - E(x))]\left[r\frac{\sigma_y}{\sigma_x}(x - \bar{x}) \right]$$

$$= br\frac{\sigma_y}{\sigma_x}E[x - E(x)^2] = \left[r\frac{\sigma_y}{\sigma_x} \right]^2 \sigma_{x^2} = r^2\sigma_{y^2}$$

$$\therefore\ r(y,\hat{y}) = \frac{r^2\sigma_{y^2}}{\sigma_y r\sigma_y} = r = r(x,y)$$

Hence the correlation coefficient between observed and estimated values of Y is the same as the correlation coefficient between X and Y.

17.9 Regression Curves:

The conditional mean $E(y \mid x = x)$) for a continuous distribution is called the regression function of Y on X and the graphs of this function of x is known as the regression of Y on X or sometimes the regression curve for the mean of Y. Geometrically, the regression function represents the Y-coordinate of the centre of mass of the bivariate probability mass in the infinitesimal vertical strip bounded by x and $x + dx$.

Similarly, the regression function of X on Y is $E(x \mid y = y)$ and the graph of this function of y is called the regression curve. (of the mean) of X on Y. In case a regression curve is a straight line, the corresponding regression is said to be linear. If one of the regression is linear, it does not however follow that the other is also linear.

Result:

Let (x,y) be a two-dimensional random variable with $E(x) = \bar{x}, E(y) = \bar{y}$, $V(x) = \sigma_{x^2}, V(y) = \sigma_{y^2}$ and let $r(x,y)$ be the correlation coefficient between X and Y. If the regression of Y on X is linear.

$$E(y \mid x) = \bar{y} + r\frac{\sigma_x}{\sigma_y}(x - \bar{x})$$

(1)

Similarly

If the regression of X and Y is linear, then

$$E(x \mid y) = \bar{x} + r\frac{\sigma_x}{\sigma_y}(y - \bar{y})$$

(2)

Proof:

Let the regression equation of Y and X be $E(y \mid x) = a + bx$

(3)

Then by the definition

$$\frac{1}{f_x(x)}\int_{-\infty}^{\infty} y\, f(y \mid x)\, dy = \int_{-\infty}^{\infty} y\, \frac{f(x,y)}{f_x(x)}\, dy$$

i.e., $$\frac{1}{f_x(x)}\int_{-\infty}^{\infty} y\, f(x,y)\, dy = a + bx$$

(4)

multiplying both sides of (4) by $f_x(x)$ and integrating w.r.to 'X' we get

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y\, f(x,y)\, dy\, dx = a\int_{-\infty}^{\infty} f_x(x)\, dx + b\int_{-\infty}^{\infty} x f_x(x)\, dx$$

$$\int_{-\infty}^{\infty} y\left[\int_{-\infty}^{\infty} f(x,y)\, dx\right] dy = a + bE(x)$$

$$\Rightarrow \int_{-\infty}^{\infty} y f_y(y)\, dy = a + bE(x)$$

$$\Rightarrow E(y) = a + bE(x) \Rightarrow \bar{y} = a + b\bar{x}$$

(5)

Multiply both sides of eq(4) by $xf_x(x)$ and integrating w.r.to 'X' , we get,

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy\, f(x,y)\, dy\, dx = a\int_{-\infty}^{\infty} x f_x(x)\, dx + b\int_{-\infty}^{\infty} x^2 f_x(x)\, dx$$

$$E(xy) = aE(x) + bE(x^2)$$

$$\Rightarrow \mu_{11} + \bar{x}\bar{y} = a\bar{x} + b(\sigma_{x^2} + \bar{x}^2) \qquad (6)$$

$$\Rightarrow \mu_{11} + \bar{x}\bar{y} = a\bar{x} + b(\sigma_{x^2} + \bar{x}^2$$

$$(\because \mu_{11} = E(xy) - E(x)E(y) = E(xy) - \bar{x}\bar{y}$$

$$\sigma_{y^2} = E(x^2) - [E(x)]^2 = E(x^2) - \bar{x}^2)$$

Solving eq (5) and (6) simultaneously, we get

$$b = \frac{\mu_{11}}{\sigma_{x^2}} \text{ and } a = \bar{y} - \frac{\mu_{11}}{\sigma_{x^2}}\bar{x}$$

Substituting in eq (3) and after simplification, we get the required equation of the line of regression of Y on X as

$$E(y \mid x) = \bar{y} + \frac{\mu_{11}}{\sigma_{X^2}}(x - \bar{x})$$

$$\Rightarrow E(y \mid x) = \bar{y} + r\frac{\sigma_y}{\sigma_x}(x - \bar{x})$$

Again by considering the line $E(x \mid y) = A + BY$ and proceeding similarly we get the equation of the line of regression of X and Y as

$$E(x \mid y) = \bar{x} + \frac{\mu_{11}}{\sigma_{Y^2}}(y - \bar{y}) = \bar{x} + r\frac{\sigma_x}{\sigma_y}(y - \bar{y})$$

17.10 Worked Out Example:

Example 1:

The following data pertain to the marks in subjects A and B in a certain examination. Mean marks A=39.5, Mean marks in B=47.5, standard deviation of marks in A=10.8, Standard deviation of marks B=16.8, Coefficient correction between marks in A and Marks in B=0.42, find the two regression lines. Find the marks in B for candidates who secured 50 marks in A.

Solution:

given mean of $A = \bar{x} = 39.5$

given mean of $B = \bar{y} = 47.5$

standard deviation of marks in $A = \sigma_x = 10.8$

standard deviation of marks in $B = \sigma_y = 16.8$

coefficient of correlation between the marks A and B $r = 0.42$

the line of regression of Y on X is

$$(y - \bar{y}) = r\frac{\sigma_y}{\sigma_x}(x - \bar{x})$$

i.e.,

$$(y - 47.5) = 0.42\frac{16.8}{10.8}(x - 39.5)$$

$$\therefore y = 0.651x + 21.82$$

(1)

The line of regression of X on Y

$$(x - \bar{x}) = r\frac{\sigma_x}{\sigma_y}(y - \bar{y})$$

i.e., $$(x - 39.5) = 0.42\frac{10.8}{16.8}(y - 47.5)$$

$$\therefore x = 0.27y + 26.67$$

(2)

When X=50 , eq(1) $\Rightarrow y = 0.651 \times 50 + 21.82 = 54.37$

Example 2:

Twenty five pairs of values of variates X and Y led to the following results N=25, $\sum X = 127$, $\sum Y = 100$, $\sum X^2 = 760$, $\sum Y^2 = 449, \sum XY = 500$. Subsequent scrutiny showed that two pairs of values were copied down as.

X	8	8
Y	14	6

X	8	6
Y	12	8

Find the correct value of r and correct line of regression

Solution: Incorrect $\sum X = 127$

The value of the incorrect values 8 and 8 , sum =16

Correct values 8 and 6 , the total value is 14

$\therefore \sum X$ should be replace by 2

The correct $\sum X = 127 - 2 = 125$

Incorrect Y value 14 and 6 then sum=20

Correct values of y 12 and 8, then sum=20

Hence there is no difference in $\sum Y$

$\therefore \text{correct} \sum Y = 100$

n=25 then $\overline{X} = \frac{\sum X_i}{n} = \frac{125}{25} = 5$

$\overline{Y} = \frac{\sum Y_i}{n} = \frac{100}{25} = 4$

incorrect $X_1^2 + X_2^2 = 8^2 + 8^2 = 64 + 64 = 128$

correct $X_1^2 + X_2^2 = 8^2 + 6^2 = 64 + 36 = 100$

$\sum X^2$ should be reduced by 28

$\therefore$ correct $\sum X^2 = 760 - 28 = 732$

incorrect $Y_1^2 + Y_2^2 = 196 + 36 = 232$

correct $Y_1^2 + Y_2^2 = 144 + 64 = 208$

Hence $\sum Y^2$ should be reduced by 24

$\therefore$ correct $\sum Y^2 = 449 - 24 = 425$

$\sigma_{X^2} = \frac{\sum X_i^2}{n} - \overline{X}^2 = \frac{732}{25} - 25 = 29.28 - 25 = 4.28$

$\sigma_X = \sqrt{4.28} = 2.07, \sigma_{Y^2} = \frac{425}{25} - 16 = 1 \text{ then } \sigma_Y = 1$

incorrect $\sum XY = X_1Y_1 + X_2Y_2 = 8 \times 14 + 8 \times 6 = 112 + 48 = 160$

correct $\sum XY = 8 \times 12 + 6 \times 8 = 96 + 48 = 144$

$\therefore \sum XY$ should be reduced by 16

Hence, correct $\sum XY = 500 - 16 = 484$

Correlation coefficient

$$r = \frac{\text{cov}(X,Y)}{\sigma_X, \sigma_Y} = \frac{\frac{1}{n}\sum XY - \overline{X}\,\overline{Y}}{\sqrt{\frac{1}{n}\sum (X_i - \overline{X})^2 \frac{1}{n}\sum (Y_i - \overline{Y}}}$$

$$= \frac{\frac{1}{n}\sum XY - \overline{X}\,\overline{Y}}{\sqrt{\left[\frac{1}{n}\sum X^2 - \overline{X}^2\right]\left[\frac{1}{n}\sum Y^2 - \overline{Y}^2\right]}}$$

$$= \frac{\frac{1}{25} \times 485 - 4 \times 5}{\sqrt{\left[\frac{0.732}{25} - 25\right]\left[\frac{425}{25} - 25\right]}} = \frac{-0.64}{2.07} = -0.31$$

The line of regression of Y on X is ,

$$Y - \overline{Y} = r\frac{\sigma_Y}{\sigma_X}\left(X - \overline{X}\right) = Y - 4 = \frac{-0.31}{2.07} \times (X - 5)$$

$$Y = -0.15X + 4.75$$

The line of regression of X on Y is ,

$$X - \overline{X} = r\frac{\sigma_X}{\sigma_Y}\left(Y - \overline{Y}\right) = X - 5 = -0.31\left[\frac{2.07}{1}\right] \times (Y - 4)$$

$$X = 0.6417Y + 7.57$$

Example 3:

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
Standard Deviation	2.1	4.2

Correlation Coefficient r=0.84

Find the two lines of regression and estimate the yield of dry bark of a plant of age 8 years.

Solution:

Given average of X, $\overline{X} = 9.2$

Given average of Y, $\overline{Y} = 16.5$

Standard deviation of X, $\sigma_X = 2.1$

Standard deviation of Y, $\sigma_Y = 4.2$

Correlation Coefficient r = 0.84

The regression Y on X is

$$Y - \overline{Y} = r\frac{\sigma_Y}{\sigma_X}\left(X - \overline{X}\right) = Y - 16.5 = 0.84\frac{4.2}{2.1}\left(X - 9.2\right)$$

$$\Rightarrow Y - 16.5 = 1.68X - 15.456 \Rightarrow Y = 1.68X + 1.044$$

The line of regression X on Y is

$$X - \overline{X} = r\frac{\sigma_X}{\sigma_Y}\left(Y - \overline{Y}\right) = X - 9.2 = 0.84\frac{2.1}{4.2}\left(Y - 16.5\right)$$

$$\Rightarrow X = 0.42Y - 6.930 + 9.2$$
$$X = 0.42Y + 2.270$$

When X=8 Years $Y = 1.68 \times 8 + 1.044 = 14.48$

$\therefore$ The yield of dry bark is 14.48

Example 4:

If x=2y+3 and y=kx+6 are the regression lines of X and Y and Y on X respectively. Then show that

(i) $0 \le k \le \frac{1}{2}$ (ii) if k=1/8 , find r and $(\overline{x}, \overline{y})$

Solution:

Given regression line x on y is x= 2y+3

Given regression line y on x is y= kx+6

(i) Then the regression coefficient of x on y is bxy =2

the regression coefficient of y on x is byx=k

$$r^2 = byx.bxy = 2k$$

We know that $0 \le k \le 1$

$$\therefore 0 \le 2k \le 1 \quad \left(\because r^2 = 2k\right)$$

$$\Rightarrow 0 \le k \le 1/2$$

(ii) If k=1/8 the regression line y on x

Y=1/8 x + 6

i.e., 8y=x+48 or 8y-x = 48 (1)

similarly, the regression line x on y

x=2y+3 or x-2y=3 (2)

solving (1) and (2) we get

-x+8y=48

x-2y=3

6y=51 and y = 51/6 = 8.5

X= 2(8.5) + 3 = 17+ 3 = 20

Then $(\bar{x}, \bar{y})$ = (20, 8.5)

$\sum \overline{X} = 130,\ \sum Y = 220, \sum X^2 = 2288\ ,\ \sum Y^2 = 5506$ and $\sum XY = 3467$

$\overline{X} = \frac{130}{10} = 13, \sum X^2 = 2288, \overline{Y} = \frac{220}{10} = 22, \sum Y^2 = 5506$

$\sigma_{x^2} = \frac{1}{n}\sum X^2 - \overline{X}^2 = \frac{2288}{10} - 169 = 59.8$

$\sigma_{x^2} = \frac{1}{n}\sum Y^2 - \overline{Y}^2 = \frac{5506}{10} - 484 = 66.6$

$S_{XY} = \sum (X_i - \overline{X})(Y_i - \overline{Y})$

$= \sum X_i Y_i - \overline{X}\sum Y_i - X_i \sum Y_i + \overline{X}\overline{Y}n$

$= 3467 - 13.220 - 22.130 + 13.22 \times 10 = 607$

$r = \frac{S_{XY}}{n\sigma_x \sigma_y} = \frac{607}{10 \times 7.8 \times 8} = 0.99$

$Y - 22 = 0.99\ \frac{8}{7.8}(X - 13)$

$Y = 1.02X + 8.7$

when price $X = 16\ , Y = 1.02 \times 16 + 8.7 = 25.02,\ r = 0.99$

$S.E = \frac{1 - r^2}{\sqrt{n}} = 0.006$

$$\therefore r = \frac{1}{2} = 0.5$$

Hence the coefficient of correlation is r=0.5

Example 5:

For any army personnel of strength 25, the regression of weight of kidney (y) on weight of heart (x) both measured n ounces is y=y-0.399x=6.934. And the regression of the weight of heart (x) on the weight of kidney (y) is x-1.224+2.46=0

Find the coefficient of correlation between X and Y and their mean values.

Solution:

The regression coefficient y on x is byx= $r\frac{\sigma_Y}{\sigma_X} = 0.399$

Since the regression line y on x is x=1.212y-2.461

$\therefore$ the regression coefficient x on y is bxy= $r\frac{\sigma_X}{\sigma_Y} = 1.212$

$\therefore r^2$=byx.bxy=0.399 * 1.212 = 0.484

$\therefore$ r =0.2

Since the two lines of regression intersect at $(\bar{x}, \bar{y})$

$\therefore$ we get $(\bar{x}, \bar{y})$ by solving the two lines of regression

i.e., y-0.399x=6.934

-0.4844+0.399x=0.982

0.576 y = 7.916

Y= 7.916/ 0.516 = 15.34

X = 1.212 (15.34) – 2.461 = 16.13

Example 6:

The equations of two regression lines obtained in a correlation analysis are 3x+12y=19, 3y+9x=46. Find (i) Coefficient of correlation (ii), Mean values off X and Y and (iii), The ratio of the coefficient of variability of X to that of Y.

Solution:

(i) Given 3x+12y=19 line of regression y on x regression coefficient of y on x is byx= $r\frac{\sigma_Y}{\sigma_X}$

i.e., 12 y =19-3x , y = 19/12 - 3/12 x

$r\frac{\sigma_Y}{\sigma_X}$ = -3/ 12 = -0.25

line of regression x on y is 3y + 9x = 46 then

9x= 46-3y = x=46/9 = 3/9 . 4

i.e., the regression coefficient x on y is

bxy= $r\frac{\sigma_X}{\sigma_Y}$ = -3/9

$\therefore r^2$=byx.bxy= -3/12 * -3/9 = 1/12

$$\therefore r = \frac{-1}{2}\sqrt{3}$$

Since both bxy and byx are negative. Therefore, r is negative

(ii) Mean values of X and Y is the point of intersection of the two lines of regression

$\therefore$ we get $(\bar{x}, \bar{y})$ by solving the equations

$3x+12y = 19$ (1)

$3y+9x=46$ (2)

(1) $x3 \Rightarrow 9x + 36y = 57$

(2) $\Rightarrow 9x + 3y = 46$

$$33y = 11$$

$$y = 11/33 = 1/3$$

$$3x + 12y = 19, \text{ i.e., } 3x + 12.\ 1/3 = 19 \Rightarrow 3x + 4 = 19$$

$$\Rightarrow 3x = 15 \Rightarrow x = 15/3 = 5$$

$$\therefore (\bar{x}, \bar{y}) = (5, y_3)$$

(iii) $$\frac{\sigma_{x^2}}{\sigma_{y^2}} = r\frac{\sigma_y}{\sigma_x} = \frac{-3}{12} = \frac{-1}{4}$$

Then

$$r^2 \frac{\sigma_{y^2}}{\sigma_{x^2}} = \frac{1}{16} \text{ and } r^2 = \frac{1}{12}$$

$$\therefore \frac{\sigma_{y^2}}{\sigma_{x^2}} = \frac{1}{16}.\frac{1}{r^2} = \frac{12}{16} = \frac{3}{4}$$

$\therefore$ The ratio of the variance of x and variance of y is 3:4

Example 7:

10 observations on place X and supply Y the following data were obtained and $\sum X = 130,\ \sum Y = 220, \sum X^2 = 2288$, $\sum Y^2 = 5506$ and $\sum XY = 3467$. Obtain the line of regression of Y on X and estimate the supply when the place is 16 units and also find the standard error of estimate.

Solution:

Hence

$$\sum \overline{X} = 130,\ \sum Y = 220, \sum X^2 = 2288,\ \sum Y^2 = 5506 \text{ and } \sum XY = 3467$$

$$\overline{X} = \frac{130}{10} = 13, \sum X^2 = 2288, \overline{Y} = \frac{220}{10} = 22, \sum Y^2 = 5506$$

$$\sigma_{x^2} = \frac{1}{n}\sum X^2 - \overline{X}^2 = \frac{2288}{10} - 169 = 59.8$$

$$\sigma_{x^2} = \frac{1}{n}\sum Y^2 - \overline{Y}^2 = \frac{5506}{10} - 484 = 66.6$$

$$S_{XY} = \sum (X_i - \overline{X})(Y_i - \overline{Y})$$

$$= \sum X_i Y_i - \overline{X}\sum Y_i - X_i \sum Y_i + \overline{X}\,\overline{Y} n$$

$$= 3467 - 13.220 - 22.130 + 13.22 \times 10 = 607$$

$$r = \frac{S_{XY}}{n\sigma_x \sigma_y} = \frac{607}{10 \times 7.8 \times 8} = 0.99$$

$$Y - 22 = 0.99\ \frac{8}{7.8}(X - 13) \Rightarrow Y = 1.02X + 8.7$$

$$\text{when price } X = 16\ , Y = 1.02 \times 16 + 8.7 = 25.02,\ r = 0.99$$

$$S.E = \frac{1 - r^2}{\sqrt{n}} = 0.006$$

Example 8:

Calculate the coefficient o correlation, also find the equation of the lines of regression and obtain an estimate of Y which should correspond on the average to X=6.2 from the following data.

X	1	2	3	4	5	6	7	8	9
Y	9	8	10	12	11	13	14	16	15

Solution:

The coefficient of correlation is $r = \frac{S_{XY}}{\sqrt{S_{XX}S_{YY}}}$ then

We have,

X_i	Y_i	$X_i = X_i - \bar{X}$	$Y_i = Y_i - \bar{Y}$	X_i^2	Y_i^2	$X_i\ Y_i$
1	9	-4	-3	16	9	12
2	8	-3	-4	9	16	12
3	10	-2	-2	4	4	4
4	12	-1	0	1	0	0
5	11	0	-1	0	1	0
6	13	1	1	1	1	1
7	14	2	2	1	4	4
8	16	3	4	4	16	12
9	15	4	3	16	9	12
45	**108**			**60**	**60**	**57**

$$\bar{X} = \frac{45}{9} = 5, \quad \bar{Y} = \frac{108}{9} = 12$$

$$\sigma_{x^2} = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n} = \frac{60}{9}, \qquad \sigma_{y^2} = \sum_{i=1}^{n} \frac{(Y_i - \bar{Y})^2}{n} = \frac{60}{9}$$

$$\therefore r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{57}{\sqrt{60 \times 60}} = \frac{57}{60} = 0.95$$

The line of regression of y on x.

$$y - \bar{} = r\frac{\sigma_y}{\sigma_x}(x - \bar{x})$$

$$(y - 12) = 0.95\frac{\sqrt{60}}{3}.\frac{3}{\sqrt{60}}(x - 5) = 0.95(x - 5)$$

$$y - 0.95x = 12 - 4.75 = 7.25 \qquad (1)$$

The line of regression of x on y is,

$$(x - 5) = 0.95(y - 12)$$

$$x - 0.95y = 5 - 11.40 = -6.4 \qquad (2)$$

when X=6.2 substitutes in (1) we get

$$y = 0.95 \times 6.2 + 7.25 = 13.14$$

Example 9:

Given $f(x,y) = x.e^{-x(y+1)}; \quad x \geq 0, y \geq 0$

Find, the regression curve of y on x

Solution:

Marginal pdf of x is given by

$$f(x,y) = \frac{f(x,y)}{f(x)} = \frac{x.e^{-x(y+1)}}{e^{-x}} = x.e^{-xy}$$

The regression curve y on x is given by

$$f(x) = \int_0^\infty f(x,y)dy = \int_0^\infty x.e^{-x(y+1)}$$

$$= x.e^{-x}\int_0^\infty e^{-xy}dy = x.e^{-x}\left|\frac{e^{-xy}}{-x}\right|_0^\infty = e^{-x}, x \geq 0$$

Conditional p.d.f of y on x is given by

$$f(x \mid y) = \frac{f(x,y)}{f(x)} = \frac{x.e^{-x(y+1)}}{e^{-x}} = x.e^{-xy}$$

The regression curve of y on x is given by

$$y = E(Y \mid X = x) = \int_0^\infty f(y \mid x)dy = \int_0^\infty y.e^{-xy}dy$$

$$y = x\left[\left|\frac{ye^{-xy}}{-x}\right|_0^\infty + \int_0^\infty \frac{e^{-xy}}{x}dy\right] = 0 + \left|\frac{e^{-xy}}{-x}\right|_0^\infty = \frac{1}{x}$$

i.e., $y = \frac{1}{x} \Rightarrow xy = 1$

which is the equation of a rectangular hyperbola. Hence, the regression of y on x is not linear.

Example 10:

In a partially destroyed laboratory record of an analysis of correlation data, the following result only are legible variance of x=9, regression equation 8x-10y+66=0, 40x-18y=214 what were

(i), the mean values of X and Y.

(ii) The correlation coefficient between X and Y and

(iii) The standard deviation of Y?

Solution:

(i) Since both the lines of regression pass through the point $(\bar{x}, \bar{y})$ we have

$8\bar{x} - 10\bar{y} + 66 = 0$ and $40\bar{x} - 18\bar{y} = 214$

Solving we get $\bar{x} = 13$, $\bar{y} = 17$

(ii) Let 8X -10Y+66=0 and 40X-18Y=214 be the lines of regression of y on x and x on y respectively. These equations can be put in the form

$$y = \frac{8}{10}x + \frac{66}{10} \text{ and } x = \frac{18}{40}y + \frac{214}{40}$$

$\therefore$ byx=Regression coefficient of y on x=$\frac{8}{10} = \frac{4}{5}$

bxy=Regression coefficient of x on y=$\frac{18}{40} = \frac{9}{20}$

Hence,

$$r^2 = byx.bxy = \frac{4}{5} \times \frac{9}{20} = \frac{9}{25}$$

$$\therefore r = \pm\frac{3}{5} = \pm 0.6$$

(iii) We have byx=$r\frac{\sigma_Y}{\sigma_X} \Rightarrow \frac{4}{5} = \frac{3}{5} \times \frac{\sigma_y}{3}$

Hence $\sigma_y = 4$

Example 11:

The following data are given for marks in English and maths in S.L.C examination of U.P.mean marks in English=39.5;mean marks in Maths=47.5 .correlation coefficient between

marks in English and maths=.42.S.D. of marks in English = 10.8 ; S.D of marks in Maths=16.8.

Forming two regressions lines calculate the acute angle between them. Also estimate the marks in maths of candidate who received 50 marks in English.

Solution:

Let x denotes the marks in English

Y denotes the marks in Maths respectively ,

given $\bar{x}$=39.5 , $\bar{y}$=47.5 , $\sigma_x = 10.8, \sigma_y = 16.8, r(x,y) = 0.42$

regression line of y on x is $y - \bar{y} = r\frac{\sigma_y}{\sigma_x}(x - \bar{x})$

or $$y - 47.5 = \frac{0.42 \times 16.8}{10.8}(x - 39.5)$$

$$\Rightarrow y = 0.65x + 21.8 \qquad (1)$$

giving its gradient m1(sat) =0.65

Similarly, the regression line of x on y is

$$(x - \bar{x}) = r\frac{\sigma_x}{\sigma_y}(y - \bar{y}) \quad \text{i.e.,} \; x = 0.274 + 26.68 \qquad (2)$$

Giving its gradient m2(say)= 1/0.27=3.7

The acute angle between them is given by

$$\tan\theta = \frac{m_1 - m_2}{1 + m_1 m_2} = \frac{0.65 - 3.7}{1 + 0.65 \times 3.7} = \frac{3.05}{3.4} = 0.9$$

$$\theta = \tan^{-1}(0.9)$$

17.11 Exercise:

1. The two regression lines are having their mean 31.6, 38 and standard deviations 3.72, 6.31 and r=0.36 find the two regression lines.

2. Variables X and Y have the join p.d.f

$$f(x,y) = \frac{1}{3}(x + y), 0 \le x \le 1, 0 \le y \le 2$$

Find (i) r(x,y) (ii) the two lines of regression (iii) The two regression curves for the means.

3.From a partially destroyed laboratory only following record could be available : x=4y+5 and y=kx+4 are the regression lines of x and Y and y on x respectively. Show that $0 \le x \le 1/4$ and if k=1/16. Find the means of two variable and the coefficient of correlation also y var x=9. Find var y and cov(x,y) when k=1/16. $\left[\text{Ans}: \sigma_y = 3/8, \text{cov}(x,y) = \frac{9}{16}\right]$

4. Obtain the two regression equations from the following data

X	1	2	3	4	5
Y	2	5	3	8	7

5. If θ is the acute angle between the two regression lines in the case of two variable x and y. Show that $\text{Tan}\theta = (1-r^2)\frac{\sigma_x\sigma_y}{r\left(\sigma_{x^2}+\sigma_{y^2}\right)}$ where r, σ_x, σ_y have their usual meanings. Interpret the results when $r = 0, r = \pm 1$.
6. If two regression coefficients are 0.8 and 0.2 what could be the value of coefficients of correlation (Ans: r=4)
7. Show that x and y are independent = the curve of regression is astraight line.
8. Is the given statement "The regression coefficients of x on y is 3.2 and that of y on x is 8" Give reasons
9. For two variables x and y with the same mean, the two regression equations are y = ax+b and $x = \alpha y + \beta$. Show that $\frac{b}{\beta} = \frac{1-\alpha}{1-\alpha}$

17.12 Summary:

In this lesson an attempt is made to explain the concepts of simple linear regression with the topics associated with them along with both theory and practical. A few examples are worked out and a good number of exercises are also given.

17.13 Technical Terms:

- Simple linear Regression
- Regression coefficients
- Angle between two regression lines
- Standard error of estimate or Residual variance
- Correlation coefficient between observed and estimated value.
- Regression curves.

17.14 Self-Assessment Questions

SHORT:

1. What is simple linear regression, and what is its basic equation?
2. Define regression coefficients and explain their role in a regression model.
3. How is the angle between two regression lines determined, and what does it indicate?
4. What is the standard error of estimate (or residual variance), and what does it measure?
5. Explain the correlation coefficient between observed and estimated values in the context of regression analysis.

ESSAY:

1. Explain the Concept and Application of Simple Linear Regression:
2. Discuss Regression Coefficients in Depth:
3. Interpreting the Angle Between Two Regression Lines:

17.5 Further Reading

(1) "Introduction to probability and statistics" by J. Susan Milton and J.C. Arnold, 4th edition, TMH (2007)

(2) "Mathematical Statistics" by R.K. Goyal, Krishna Prakashan Media (P) Ltd, Meerut.

(3) "Fundamentals of Mathematical Statistics" by S.C. Gupta and V.K. Kapoor, S.Chand & Sons, New Delhi

Dr. B. Sri Ram

LESSON-18

CORRELATION

AIMS AND OBJECTIVES:

This lesson is prepared in such a way that after studying this lesson the student expected to have a clear comprehension of the theory and practical utility about the concept of correlation properties and applications, which are the important areas of investigation and statistical data analysis. The students will be having and well equipped with both theoretical as well as practical aspects of correlation.

STRUCTURE OF THE LESSON:

18.1 Introduction
18.2 Notion of correlation
18.3 Karl Pearson's coefficient of correlation
18.4 Properties of Correlation coefficient
18.5 Assumptions underlying Karl Pearson's correlation coefficient
18.6 Scatter diagram
18.7 Worked out examples
18.8 Exercises
18.9 Summary
18.10 Technical Terms
18.11 Self-Assessment Questions
18 .12 Further Reading

18.1 INTRODUCTION:

The distribution of people according to heights or weights, annual rainfall in a certain area, the agricultural yield and so on have been studied. Now, if we measure both heights and weights to find some sort of relation between the two-whether tall persons are heavier than short ones, an increase in rainfall results in greater agricultural yield, more consumption of rice results in increase of both rate and so on. We may find that a change in one variable results in a direct or inverse change in the other or does not have any effect on the second variable. The relationship between two variables such that change in the other and also greater change in one variable results in a corresponding greater change in the other is known

as correlation. If the second variable is unaffected by a change in the first example the heights of the fathers and marks in mathematics and sons, they are said to be statistically independent.

Regression technique provides the actual relationship between two or more variables. But scientists are not always interested in this linear or curvilinear relationship. Often the interest lies only in knowing the extent of interdependence between two or more variables. In this situation, correlation methods serve our purpose. If the two variables say x and y are linearly related, they are said to be correlated. The correlation between two variables is also known as simple correlation. The measure of correlation was given by Prof. Karl Pearson in 1896 in the form of correlation coefficient.

18.2 NOTION OF CORRELATION:

The relationship between the two variables such that a change in one variable results in a positive or negative change in the other and also greater change in one variable results in a corresponding greater change in the other is called a correlation.

For a change in one variable, there is a corresponding change in the other variable. The variables are said to be correlated

a. If the two variables deviate in the same direction, the correlation is said to be direct or positive.

b. If the variables deviate in the opposite direction, the correlation is said to be inverse or negative.

c. If the change in one variable corresponds to a proportional change in the other variable then the correlation is said to be perfect.

18.3 KARL PEARSON COEFFICIENT OF CORRELATION

Measure of intensity or degree of linear relationship between two variables is called correlation coefficient and it is denoted by 'r' or r(x,y)

Coefficient of correlation r=Coefficient of correlation $r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$

Coefficient of correlation $r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$

Where $S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$

$S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2$

$S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$

$\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ $\sigma_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$

$\text{cov}(xy) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \Rightarrow \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = n\,\text{cov}(xy)$

$\sum_{i=1}^{n}(x_i - \bar{x})^2 = n\sigma_x^2 \Rightarrow \sqrt{\sum_{i=1}^{n}(x_i - \bar{x})2} = \sigma_x \cdot\sqrt{n}$

$\sum_{i=1}^{n}(y_i - \bar{y})^2 = n\sigma_y^2 \Rightarrow \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})2} = \sigma_y \cdot\sqrt{n}$

Therefore, r can be written as $r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n.\sigma_x\,\sigma_y}$

$\Rightarrow r = \frac{n\text{cov}(x,y)}{n.\sigma_x\,\sigma_y} = \frac{\text{cov}(x,y)}{\sigma_x\,\sigma_y}$

18.4 PROPERTIES OF COEFFICIENT OF CORRELATION:

1. The coefficient of correlation lies between -1 and +1.

We have $r(x,y) = \frac{\text{cov}(xy)}{\sigma_x\,\sigma_y} = \frac{\frac{1}{n}\sum_i(x_i - \bar{x})(y_i - \bar{y})}{[\frac{1}{n}\sum(x_i - \bar{x})2\,\frac{1}{n}\sum(y_i - \bar{y})2\,]1/2}$

$i = 1,2\ldots..n$

$\therefore r^2(x,y) = \frac{[\sum a_i b_i]2}{[\sum a_i^2\,][\sum b_i^2\,]}$, where $a_{i=x_i - \bar{x}}$, $b_{i=y_i - \bar{y}}\}$ $\rightarrow(1)$

From the Schwartz inequality which states that if a_i, b_i for i=1,2,…n are real quantities then

$$(\sum_{i=1}^{n} a_i b_i)^2 \leq (\sum_{i=1}^{n} a_i^2)(\sum_{i=1}^{n} b_i^2)$$

The sign of equality holding if and only if

$$\frac{a_1}{b_1} = \frac{a_2}{b_2} = \cdots\cdots = \frac{a_n}{b_n}$$

Using Schwartz inequality, we get from (1)

$$r^2(x,y) \leq 1 \quad \text{i.e, } |r(x,y)| \leq 1 \quad \Longrightarrow -1 \leq r(x,y) \leq +1.$$

Hence correlation coefficient cannot exceed unity numerically. It always lies between +1 and -1. If r =+1, the correlation is perfect and positive and if r =-1 correlation is perfect and negative.

2. Correlation coefficient is independent of origin and scale.

We shall prove that $r(x,y)=r(u,v)$

Let $u=\frac{x-a}{h}$ so that $x=a+uh$

$v=\frac{y-b}{k}$ so that $y=b+kv$.

Where a, b, h, k are constants. Since x=a+uh and y=b+kv on taking expectations we get

$E(x)=E[a+uh]=a+hE(u) \Longrightarrow x-E(x)=a+uh-a-hE(u)=h[u-E(u)]$

$E(y)=E[b+kv]=b+kE(v) \Longrightarrow y-E(y)=b+kv-b-kE(v)=k[v-E(v)]$

Then

$cov(x,y) = E[(x-E(x)(y-E(y)]$

$=E[h(u-E(u))k(v-E(v))]$

$=hk.[E(u-E(u)).(v-E(v))]$

$=hk.cov(u,v).$ $\rightarrow(1)$

$\sigma_x^2=E[(x-E(x))^2]=E[h^2(u-E(u))^2]=h^2.\sigma_u^2 \Longrightarrow \sigma_x=h.\sigma_u (h>0)$ $\rightarrow(2)$

$\sigma_y^2=E[(y-E(y))^2]=E[k^2(v-E(v))^2]=k^2.\sigma_v^2 \Longrightarrow \sigma_y=k.\sigma_v (k>0)$ $\rightarrow(3)$

Also we know that $r(x,y)=\frac{cov(xy)}{\sigma_x \sigma_y}=\frac{hkcov(u,v)}{h\sigma_u .k\sigma_y}$

$$r(x,y)=\frac{cov(xy)}{\sigma_{u.}\sigma_v} \quad \rightarrow(4)$$

Two independent variables are uncorrelated

We know that from definition of correlation

$$r(x,y)=\frac{cov(xy)}{\sigma_{x.}\sigma_y} \quad \rightarrow(1)$$

Where $cov(x,y)=E[(x-E(x))(y-E(y))]$

$=E[xy+E(x).E(y)-xE(y)-E(x)y]$

$=E(xy)+E(x).E(y)-E(x).E(y)-E(x).E(y)]$

$cov(x,y)=E(xy)-E(x)E(y) \quad \rightarrow(2)$

If x and y are independent variables, we write

$E(xy)=E(x).E(y) \quad \rightarrow(3)$

$\therefore cov(x,y)=E(x).E(y)-E(x).E(y)=0$ [$\therefore$ from using (3) in (2)]

$cov(x,y)=0 \quad \rightarrow(4)$

Hence, if x and y are independent variables from equation(4)

equation(1) $\Rightarrow r(x,y)=\frac{cov(xy)}{\sigma_{x.}\sigma_y}=\frac{0}{\sigma_{x.}\sigma_y}=0$

$\therefore r(x,y)=0$

i.e., The two independent variables are uncorrelated.

Note 1:

The converse of the above property is not true.

i.e., two uncorrelated variables may not be(true)independent.

For example if $x\sim N(0,1)$ and $y=x^2$

Since x~N(0,1), $E(x)=0=E(x^3)$ then

$cov(x,y)=E(xy)-E(x).E(y)$

$=E(x.x^2)-E(x).E(y)$ $(\therefore y=x^2)$

$=E(x^3)-E(x)E(y)=0-0.E(y)=0$

$\therefore E(x)=0, E(x^3)=0$

Hence $r(x,y)=0$, therefore we conclude that x and y are uncorrelated but not independent.

Note 2:

Let x be random variable with p.d.f

$f(x)=1/2$, $-1\leq x\leq 1$

Let $y=x^2$ Here we shall get

$E(x)=0$ and $E(xy)=E(x^3)=0\Longrightarrow r(x,y)=0$

Note 3:

Also, the converse of the above result holds in the following cases:

(a) If x and y are jointly normally distributed with r(x,y)=0 then they are independent. If r=0 then

$f(x,y)=\frac{1}{\sigma_{x.\sqrt{2\pi}}}\exp[\frac{-1}{2}[\frac{x-\mu_x}{\sigma_x}]^2].\frac{1}{\sigma_{y.\sqrt{2\pi}}}\exp[\frac{-1}{2}[\frac{y-\mu_y}{\sigma_y}]^2]$

$\therefore f(x,y)=f_1(x).f_2(y)$

$\Longrightarrow$x and y are independent.

(b) If each of the two variables x and y takes two values 0,1 with positive probabilities, then $r(x,y)=0\Longrightarrow$x and y are independent.

Proof: Let x take the values 1 and 0 with positive probabilities p_1 and q_1, respectively and let y take the values 1 and 0 with possibilities p_2 and q_2 respectively then,

$r(x,y)=0\Longrightarrow cov(x,y)=0$

$\Longrightarrow 0=E(xy)-E(x)E(y)$

$=1-p(x=1\cap y=1)-[1.p(x=1).1.p(y=1)]$

$=p(x=1\cap y=1)-p_1p_2$

$\Longrightarrow p(x=1\cap y=1)=p_1p_2=p(x=1).p(y=1)$

$\Longrightarrow$x and y are independent.

18.5 Assumptions underlying Karl Pearson's correlation coefficient

Pearsonian correlation coefficient r is based on the following assumptions.

(i) The variables x and y under study are linearly correlated. i.e when the data plotted on a graph, the scatter diagram of the data will give a straight line curves.

(ii) Each of the variables is being affected by a large number independent contributory causes of such a nature as to produce normal distribution. For example, the variables(series)relating to ages, heights, weights, supply, price etc. confirm to this assumption.

(iii)The forces so operating on each of variables series are not independent of each other but are related in casual fashion. In other word, cause and effect relationship exists between different form on the items of the two variables series. These forces must be common to both the series. If the operating forces are entirely independent of each other and not related in any fashion, then there cannot be any correlation between the variables under study.

For example, the correlation coefficient between

(a) The series of heights and income of individuals over a period of time.

(b) The series of marriage rate and the rate of agricultural production in a country over a period of time.

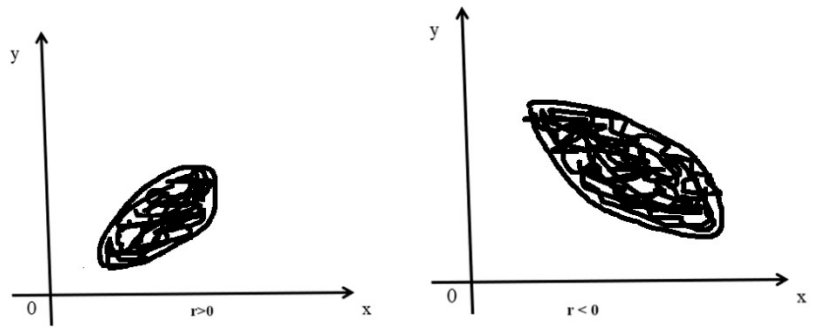
(c) The series relating to the size of the shoe and intelligence of group of individuals,.

Should be zero, since the forces affecting the two variables series in each of the above cases are entirely independent of each other.

However, if any of the above cases the value of 'r' for given set of data is not zero, then such correlation is termed as chance correlation or spurious or non-sense correlation.

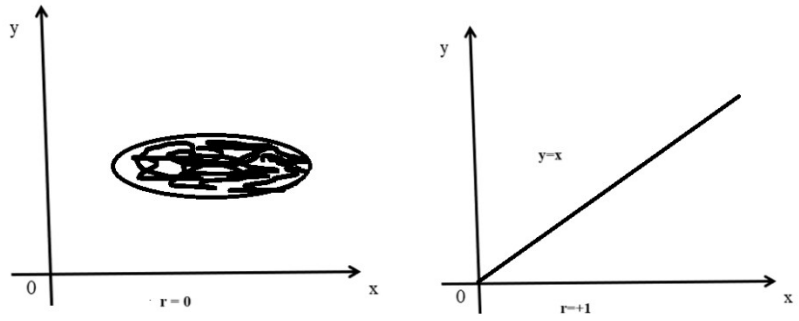
18.6 SCATTER DIAGRAM:

It is the simplest way of the diagrammatic representation of bivariate data. Thus for the bivariate distribution (x_i,y_i) i=1,2,…n if the values of the variables x and y be plotted along the x-axis and y-axis respectively. In the xy-plane, the diagram of the dots so obtained is known as scatter diagram. From the scattering diagram can form a fairly good, though vague, idea whether the variables are correlated or not, for example, if the points are very dense i.e very close to each other we should expect a fairly good amount of correlation between the variables and if the points are widely scattered, a poor correlation is expected. But this method, however is not suitable if the number of observations fairly large. Following are the figures of the scattered data r>0, r<, r=0, r=±1.



(a) (b)

Fig 18.1 (a) r >0 (b) r<0



(a) (b)

Fig 18.2 (a) r = 0 (b) r=+1

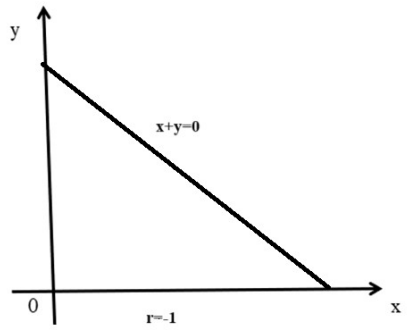


Fig 18.3 r=-1

18.7 WORKED OUT EXAMPLES:

Example 1: The variables x and y are connected by the equation ax+by+c=0 show that the correlation between them is -1 if the sign of a and b are all alike and +1 if they are different

Solution: Given ax+by+c=0 $\longrightarrow(1)$

Taking expectations on both sides of equation(1)

$E(ax+by+c)=E(0) \Longrightarrow aE(x)+bE(y)+c=0 \longrightarrow(2)$

$[\therefore E(0)=0, E(c)=c]$

$\therefore$ from (1) and (2) we get

$ax+by+c-[aE(x)+bE(y)+c]=0$

$\Longrightarrow a[x-E(x)]+b[y-E(y)]=0$

$\Longrightarrow a[x-E(x)=-b[y-E(y)] \Longrightarrow [x-E(x)]=-\frac{-b}{a}[y-E(y)] \longrightarrow(3)$

But

$cov(x,y)=E[x-E(x)(y-E(y)]=E[\frac{-b}{a}[y-E(y)][y-E(y)]$From equation(3)

$cov(x,y)=E[\frac{-b}{a}]E[y-E(y))^2]$

$=\frac{-b}{a}\sigma_y^2 \quad (\therefore \sigma_y^2 = E[(y-E(y))^2]) \longrightarrow(4)$

Taking expectation and squaring on both sides of equation(3)

We get

$\sigma_x^2 = E[x-E(x)]^2 = \frac{b^2}{a^2} E[y-E(y)]^2 = \frac{b^2}{a^2}\sigma_y^2$ $\longrightarrow(5)$

Correlation $r(x,y) = \frac{cov(xy)}{\sqrt{\sigma_x^2}\sqrt{\sigma_y^2}}$

$$= \frac{-b/a\ \sigma_y^2}{\sqrt{\frac{b^2}{a^2}\sigma_y^2}\sqrt{\sigma_y^2}} = \frac{-b/a\ \sigma_y^2}{|b/a|\ \sigma_y^2}$$

[$\therefore$using equations (4),(5)]

$$= \begin{cases} +1, \text{if b and a are of opposite sign.} \\ -1, \text{if b and a are of same sign.} \end{cases}$$

Hence, the correlation between them is -1 if the sign of a and b are alike and +1 if they are different.

Example 2:

If x and y are two random variables with variances σ_x^2 and σ_y^2 respectively and r is the coefficient of correlation between them. If U=x+ky and V= $x+\frac{\sigma_x}{\sigma_y}y$, Find the value of k so that U and V are uncorrelated.

Solution: Given $U=x+ky$ and $V=x+\frac{\sigma_x}{\sigma_y}y$ $\longrightarrow(1)$

Taking expectations on both sides of equation (1), we get

$E(U)=E(x)+k.E(y)$ and $E(V)=E(x)+\frac{\sigma_x}{\sigma_y}E(y)$ $\longrightarrow(2)$

Then $U-E(U)=x+ky-E(x)-k.E(y)=[x-E(x)]+k[y-E(y)]$ $\longrightarrow(3)$

And $V-E(V)=x+\frac{\sigma_x}{\sigma_y}y-E(x)-\frac{\sigma_x}{\sigma_y}E(y)=[x-E(x)]+\frac{\sigma_x}{\sigma_y}[y-E(y)]$ $\longrightarrow(4)$

$cov(u,v)=E[[U-E(U][V-E(V)]]$

$=E[[[x-E(x)]+k[y-E(y)][x-E(x)]+\frac{\sigma_x}{\sigma_y}[y-E(y)]]$

[From (3) and (4) equations]

$=\sigma_x^2 x + \frac{\sigma_x}{\sigma_y}\text{cov}(x,y)+k\text{cov}(x,y)+k.\frac{\sigma_x}{\sigma_y}.\sigma_y^2$

$=[\sigma_x^2 x+k\sigma_x\sigma_y]+[\frac{\sigma_x}{\sigma_y}.+k]\text{cov}(x,y)$

$=\sigma_x[\sigma_x+k\sigma_y]+[\frac{\sigma_x+k\sigma_y}{\sigma_y}]\text{cov}(x,y)$

$\text{cov}(u,v)= [\sigma_x+k\sigma_y][\sigma_x+\frac{\text{cov}(x,y)}{\sigma_y}]= (\sigma_x+k\sigma_y)(1+r)\sigma_x$

$[\therefore\frac{\text{cov}(x,y)}{\sigma_x\sigma_y}=r\Longrightarrow\frac{\text{cov}(x,y)}{\sigma_x}=r.\sigma_x]$

u and v are uncorrelated if

$r(u,v)=0\Longrightarrow \text{cov}(u,v)=0$

i.e if $(\sigma_x+k\sigma_y)(1+r)\sigma_x=0$

$\Longrightarrow \sigma_x+k\sigma_y=0 \qquad (\therefore\sigma_x\neq 0, r\neq -1)$

$\Longrightarrow k=-\frac{\sigma_x}{\sigma_y}$

Hence the value of k is $-\frac{\sigma_x}{\sigma_y}$

Example 3:

The random variables x and y are jointly normally distributed, and U and V are defined by $U=x\cos\alpha+y\sin\alpha$, $V=y\cos\alpha-x\sin\alpha$

Show that U and V will be uncorrelated if

$\text{Tan}2\alpha=\frac{2r\sigma_x\sigma_y}{\sigma_x^2-\sigma_y^2}$

Where r =corr(x,y), $\sigma_x^2=\text{var}(x)$ and $\sigma_y^2=\text{var}(y)$.

Are U and V are independent?

Solution: We have cov(u,v)=E[(u-E(U))(v-E(V)]

$=E[(x-E(x)\cos\alpha+(y-E(y)\sin\alpha].[(y-E(y)\cos\alpha-[x-E(x)]\sin\alpha]$

$=\cos^2\alpha\text{cov}(x,y)-\sin\alpha\cos\alpha.\sigma_x^2+\sin\alpha\cos\alpha\sigma_y^2-\sin^2\alpha.\text{cov}(x,y)$

$=(\cos^2\alpha-\sin^2\alpha)\text{cov}(x,y)-\sin\alpha\cos\alpha(\sigma_x^2-\sigma_y^2)$

$=\cos2\alpha.\text{cov}(x,y)-\sin\alpha\cos\alpha(\sigma_x^2-\sigma_y^2)$ $(\therefore \cos^2\alpha-\sin^2\alpha=\cos2\alpha)$

u and v will be uncorrelated if and only if

r(u,v)=0, i.e., if cov(u,v)=0

i.e., if $\cos2\alpha\text{cov}(x,y)-\sin\alpha\cos\alpha(\sigma_x^2-\sigma_y^2)=0$

or if $\cos2\alpha\, r\sigma_x\sigma_y=\frac{\sin2\alpha}{\sigma_y}(\sigma_x^2-\sigma_y^2)$

$(\therefore\frac{\text{cov}(x,y)}{\sigma_x\sigma_y}=r\Rightarrow\text{cov}(x,y)=r\,\sigma_x\,\sigma_y.2\sin\alpha\cos\alpha=\sin2\alpha)$.

Or if $\frac{\sin2\alpha}{\cos2\alpha}=\frac{2\,r\,\sigma_x\,\sigma_y}{(\sigma_x^2-\sigma_y^2)}\Rightarrow\text{Tan}2\alpha=\frac{2\,r\,\sigma_x\,\sigma_y}{(\sigma_x^2-\sigma_y^2)}$ $(\therefore\text{Tan}2\alpha=\frac{\sin2\alpha}{\cos2\alpha})$

However, r(u,v)=0 does not imply that the variables U and V are independent.

Example 4:

Calculate the correlation coefficient for the following heights (in inches) of fathers(X) and their sons(Y):

X 65 66 67 67 68 69 70 72

Y 67 68 65 68 72 72 69 71

Solution: Coefficient of correlation is given by $r(x,y)=\frac{\text{cov}(xy)}{\sigma_{x.}\sigma_y}$ $\rightarrow(1)$

Where $\text{cov}(x,y)=\frac{1}{n}\sum XY-\overline{X}\,\overline{Y}$, $\boldsymbol{\sigma}_x=\sqrt{[\frac{1}{n}\sum X^2-\overline{X^2}]}$, $\boldsymbol{\sigma}_y=\sqrt{[\frac{1}{n}\sum Y^2-\overline{Y^2}]}$

Calculations for correlation coefficient

	X	Y	X^2	Y^2	XY	
	65	67	4225	4489	4355	
	66	68	4356	4624	4488	
	67	65	4489	4225	4355	
	67	68	4489	4624	4556	
	68	72	4624	5184	4896	

	69	72	4761	5184	4968	
	70	69	4900	4761	4830	
	72	71	5184	5041	5112	
Total	544	552	37028	38132	37560	

$$\bar{X} = \frac{1}{n}\sum X = \frac{544}{8} = 68\ ,\ \bar{Y} = \frac{1}{n}\sum Y = \frac{552}{8} = 69$$

$$r(x,y) = \frac{\mathrm{cov}(xy)}{\sigma_x \sigma_y} = \frac{\frac{1}{n}\sum XY - \bar{X}\bar{Y}}{\sqrt{[\frac{1}{n}\sum X^2 - \bar{X^2}]}\sqrt{[\frac{1}{n}\sum Y^2 - \bar{Y^2}]}}$$

$$r(x,y) = \frac{\frac{1}{8}\times 37560 - 68\times 69}{\sqrt{[\frac{37028}{8} - (68)^2]}\sqrt{[\frac{38132}{8} - (69)^2]}}$$

$$r(x,y) = \frac{4695 - 4692}{\sqrt{[4628.5 - 4624][4766.5 - 4761]}}$$

$$= \frac{3}{\sqrt{4.5\times 5.5}} = 0.603$$

Hence the correlation coefficient between x,y for the given data is r(x,y)=0.603.

Example 5:

A computer while calculating correlation coefficient between two variables x and y from 25 pairs of observations obtained the following results:

n=25, $\sum X=125$, $\sum X^2=650$, $\sum Y=100$, $\sum Y^2=460$, $\sum XY=508$

It was later discovered at the time of checking that he had copied down two pairs as

X	6	8
Y	14	6

While the connected values were

X	8	6
Y	12	8

obtain the correct value of correlation coefficient.

Solution:

Corrected $\sum X=125-6-8+8+6=125$

Corrected $\sum Y=100-14-6+12+8=100$

Corrected $\sum X^2=650-6^2-8^2+8^2+6^2=650$

Corrected $\sum Y^2=460-14^2-6^2+12^2+8^2=650$

Corrected $\sum XY=508-6\times 114-8\times 6+8\times 12+6\times 8=520$

$\bar{X}=\frac{1}{n}\sum X=\frac{1}{25}\times 125=5,\ \bar{Y}=\frac{1}{n}\sum Y=\frac{1}{25}\times 100=4.$

$\text{cov}(x,y)=\frac{1}{n}\sum XY-\bar{X}\ \bar{Y}=\frac{1}{25}\times 520-5\times 4=4/5.$

$\sigma_x^2=\frac{1}{n}\sum X^2-\bar{X}^2=\frac{1}{25}\times 650-(5)^2=1$

$\sigma_y^2=\frac{1}{n}\sum Y^2-\bar{Y}^2=\frac{1}{25}\times 436-16=\frac{36}{25}$

$\therefore$ Corrected $r(x,y)=\frac{\text{cov}(xy)}{\sigma_{x.}\sigma_y}=\frac{4/5}{1\times 6/5}=2/3=0.67$

Hence the corrected correlation coefficient for the given data is r(x,y)=0.67.

Example 6:

Show that if x^1, y^1 are the deviations of the random variables x and y from their respective means then

$$\text{(i) } r=1-\frac{1}{2N}\sum_i\left[\frac{x_i^1}{\sigma_X}-\frac{y_i^1}{\sigma_y}\right]^2$$

(ii)r= $-1+\frac{1}{2N}\sum_i\left[\frac{x_i^1}{\sigma_X}-\frac{y_i^1}{\sigma_y}\right]^2$

Deduce that $-1\le r\le+1$

Solution: Here $x_i^1=(x_i-\bar{x})$ and $y_i^1=(y_i-\bar{y})$

R.H.S=$1-\frac{1}{2N}\sum_i\left[\frac{x_i^1}{\sigma_X}-\frac{y_i^1}{\sigma_y}\right]^2$

$$=1-\frac{1}{2N}\sum_i\left[\frac{x^1{}_i^2}{\sigma_x^2}+\frac{y^1{}_i^2}{\sigma_y^2}-\frac{2x^1{}_iy^1{}_i}{\sigma_X\sigma_y}\right]$$

$$=1-\frac{1}{2N}\left[\frac{1}{\sigma_x^2}\sum_i x^1{}_i^2+\frac{1}{\sigma_y^2}\sum_i y^1{}_i^2-\frac{2}{\sigma_x\sigma_y}\sum_i x_i^1y_i^1\right]$$

$$=1-\frac{1}{2N}\left[\frac{1}{\sigma_x^2}\sum_i(x_i-\bar{x})^2+\frac{1}{\sigma_y^2}\sum_i(y_i-\bar{y})^2-\frac{2}{\sigma_x\sigma_y}\sum_i(x_i-\bar{x})(y_i-\bar{y})\right]$$

$$=1-\frac{1}{2}\left[\frac{1}{\sigma_x^2}.\sigma_x^2+\frac{1}{\sigma_y^2}.\sigma_y^2-\frac{2}{\sigma_x\sigma_y}r\sigma_x\sigma_y\right]$$

$$=1-\frac{1}{2}[1+1-2r]=r$$

(ii)Proceeding in similar manner, we will get

R.H.S= $-1+\frac{1}{2}[1+1+2r]=r$

Deduction: Since $\left[\frac{x_i^1}{\sigma_X}\pm\frac{y_i^1}{\sigma_y}\right]^2$ being the square of real quantity is always non-negative $\sum_i\left[\frac{x_i^1}{\sigma_X}\pm\frac{y_i^1}{\sigma_y}\right]^2$ is also non-negative. From part(i), we get r=1-(some non-negative quantity)

$\Rightarrow r\le 1 \quad \rightarrow(1)$

Also from part(ii) we get

r=-1+(some non-negative quantity) $\Rightarrow$-1≤r $\rightarrow$(2)

The sign of equality in (1) and (2) holds if and only if x_i^1

$\frac{x_i^1}{\sigma_X} - \frac{y_i^1}{\sigma_y} = 0$, $\frac{x_i^1}{\sigma_X} + \frac{y_i^1}{\sigma_y} = 0$, $\forall$ i=1,2,….n respectively.

From (1) and (2), we get -1≤r≤1.

18.8 Exercises

1. Calculate the coefficient of correlation between x and y for the following

X 1 3 4 5 7 8 10

Y 2 6 8 10 14 16 20 **[Ans: r(x,y)=±1]**

2(a) From the following data, complete the coefficient of correlation between x and y.

	X-series	Y-series
No.of items	15	15
Arithmetic Mean	25	18
Sum of squares of deviations from mean	136	138

Summation of product deviations of x and y series from the respective arithmetic means=122.[Ans:r(x,y)=0.891]

(b) coefficient of correlation between two variables x and y is 0.32. Their covariance is 7.86. The variance of x is 10. Find the standard deviation of y-series.

(c) In two sets of variables x and y with 50 observations each, the following data were observed.

$\overline{x}$=10, $\sigma_x = 3$,$\overline{y}$=6, σ_y=2 and r(x,y)=0.3.

But on subsequent verification it was found that one value of x(=10) and the value of y(=6) were inaccurate and hence weeded out. With the remaining parts of values how is the original value of r affected?

3. If x_1 and x_2 are two random variables with means μ_1 and μ_2, variances σ_1^2,σ_2^2 and correlation coefficient r, find the correlation coefficient between $U=a_1x_1+a_2x_2$ and $V=b_1x_1+b_2x_2$

Where a_1,a_2,b_1,b_2 are constants.

4. If U=ax+by and V=bx-ay, where x and y are measures from their respective means and if U and V are uncorrelated, r the coefficient of correlation between x and y is given by the equation

$$\sigma_u\sigma_v=(a^2+b^2)\sigma_x\sigma_y(1-r^2)^{1/2}.$$

5. A coin is tossed n times. If x and y denote the (random) number of heads and number of trails turned up respectively, show that r(x,y)=-1.

6. Two dice are thrown, their scores being a and b. The first die is left on the table while the second is picked up and thrown again giving the scores c: Suppose the process is repeated a large number of times. What is the correlation coefficient between X=a+b and Y=a+c? [Ans:r(x,y)=1/2]

7. If x and y are independent random variables with means μ_1 and μ_2 and variances σ_1^2,σ_2^2respectively, show thatthe correlation coefficient between U=x and V=x-y in terms of μ_1 , $\mu_2\sigma_1^2$ andσ_2^2 is $\sigma_1 \Big/ \sqrt{\sigma_1^2+\sigma_2^2}$.

8. x1 and x2 are independent variables with means 5 and 10 and standard deviations 2 and 3 respectively. Obtain r(u,v) when $U=3x_1+4x_2$ and $V=3x_1-x_2$[Ans:0]

9. A prognostic test in mathematics was given to 10 students who were about to begin a course in Statistics. The scores(x) in their test were examined in relations to scores(y) in the final examination in statistics. The following results were obtained.

$\sum x=71,\sum y=70,\sum x^2=555,\sum y^2=526$,and $\sum xy=527$

Find the coefficient of correlation between x and y.

10. Using the formula $\sigma_{x-y}^2=\sigma_x^2+\sigma_y^2-2r(x,y)\sigma_x\sigma_y$.

Obtain the correlation coefficient between the heights of fathers(X) and of the sons(Y) from the following data.

X:	65	66	67	68	69	70	71	67
Y:	67	68	64	72	70	67	70	68

18.9 Summary:

In this lesson an attempt is made to explain the concepts of correlation associated with them along with both theory and practical. A few examples are worked out and a good number of exercise problems are given.

18.10 Technical Terms:

- Correlation
- Karl Pearson’s coefficient of correlation and Assumptions
- Scatter diagram

18.11 Self-Assessment Questions

SHORT:

1. What is correlation, and why is it important in statistical analysis?
2. Define Karl Pearson's coefficient of correlation.
3. List two key properties of the correlation coefficient.
4. What are the main assumptions underlying Karl Pearson’s correlation coefficient?
5. What is a scatter diagram, and how is it used in correlation analysis?

ESSAY:

1. explain the notion of correlation and discuss how Karl Pearson’s coefficient measures the linear relationship between two variables.
2. Discuss the properties of the correlation coefficient in detail.
3. Examine the assumptions underlying Karl Pearson’s correlation coefficient and explain why these assumptions are crucial.
4. Describe the construction and interpretation of a scatter diagram in the context of correlation analysis.
5. Critically compare Pearson's correlation coefficient with other measures of association.

18 .12 Further Reading

(1) "Introduction to probability and statistics" by J. Susan Milton and J.C. Arnold, 4th edition, TMH (2007)

(2) "Mathematical Statistics" by R.K. Goyal, Krishna Prakashan Media (P) Ltd, Meerut.

(3) "Fundamentals of Mathematical Statistics" by S.C. Gupta and V.K. Kapoor, S.Chand & Sons, New Delhi

Dr. B. Sri Ram

LESSON - 19

MULTIPLE LINEAR REGRESSION MODELS - LEAST SQUARE PROCEDURES FOR MODEL FITTING A MATRIX NOTATION

OBJECTIVE:

This Lesson is prepared in such a way that after studying the material the student is expected to have a through comprehension of the concept "Least-Square Procedures for model fitting and matrix notation" are the pivotal of statistical inference and analysis. The student would be equipped with theoretical as well as practical aspects of concepts.

STRUCTURE OF THE LESSON:

19.1 Introduction

19.2 Multiple Linear Regression Models-Least Square Procedure for Model Fitting

19.3 Multiple Linear Regression Models-A Matrix Approach to Least Squares

19.4 Worked out Examples

19.5 Exercises

19.6 Summary

19.7 Technical Terms.

19.8 Self assessment questions

19.9 Further Reading

19.1 INTRODUCTION:

When two variables are under study simultaneously is the study of regression and correlation. But scientific, social and economic phenomena do not confined to two variables only. Many studies involve more than two variables. In these studies, we often need to give an actual relationship between three or more variables and/ or to explain the strength of association between them. For this, Multiple/Multivariate regression are strong tools. For instance, the cost of production of a manufactured product mainly depends on the cost of raw materials, The labour charges and the cost

of energy. The cost of a crop mainly depends mainly upon the cost of Seeds, Fertilizers, Irrigation, Pesticides and many farm operations. In both the examples, the cost of the produced product is a dependent factor, While others are independent factors. If we want to establish the relationship between the dependent variable and independent variables, A mathematical equation can be given to do this. This type of mathematical equation is known as a mathematical model. The equation pertaining to such a relationship maybe of any type. But we will only deal with a linear relationship which represents a plane or hyper plane according to the number of variables involved. Fitting of a regression equation means, the estimation of parameters involved in the models. A mathematical model with the dependent variable Y and K independent variables $x_1, x_2 .....x_k$ uis

$$y = a_0 + a_1 x_1 + a_2 x_2 + ... + a_k x_k + \varepsilon \text{ s}$$

This type of regression equation is also known as multiple regression equation or the prediction equation, with y as predictant and $x_1, x_2 .....x_k$ are predictors. ε is the error which is distributed normally with mean 0 and variance σ^2 i.e, $\varepsilon \sim N(0, \sigma^2)$. In this lesson we discuss the MLR models-Least Square Procedure for Model Fitting, A matrix approach to least squares with respective applications.

19.2 MULTIPLE LINEAR REGRESSION MODELS:

Least squares procedures for model fitting:

Having collected some data, it is desirable to find out the form of universe of which the observed values are regarded as a sample. In other words, we try to find a functional relationship between the observed values so as to have a clear picture of the universe of which our observations are a part. It is neither necessary nor possible that all the observed values should strictly satisfy this relationship, but the curve, representing this relationship, should as far as possible pass closely to all the points, the difference between the observed values and expected values is known as residual and the task is to minimize these residuals. Since, these differences may be positive in some cases and negative in others; it is more convenient to make the sum of squares of these residuals a minimum . This is known as the method of least squares.

Fitting of a straight line:

Let us consider the fitting of a straight line

$$y = a + bx \text{ ---------(1)}$$

to a set of n points $(x_{i,} y_i)\,; i = 1,2,3.....n$. Equation (1) represents a family of straight lines for different values of the orbitrary constants 'a' and 'b'. The problem is to determine 'a' and 'b' so that the line one is the line of best fit. According to the principle of least squares we have to determine a and b so that

$$E = \sum_{i=1}^{n} \left(y_i - a - bx_i \right)^2$$

is minimum. From the principle of maxima and minima the partial derivatives of E, (w.r.to) a and b should vanish separately. i.e..,

$$\left.\begin{aligned} \frac{\partial E}{\partial a} = 0 \Rightarrow -2\sum_{i=1}^{n} (y_i - a - bx_i) = 0 \\ \frac{\partial E}{\partial b} = 0 \Rightarrow -2\sum_{i=1}^{n} x_i (y_i - a - bx_i) = 0 \end{aligned}\right\} \text{----------(2)}$$

$$\left.\begin{aligned} \Rightarrow \sum_{i=1}^{n} y_i = na + b\sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i y_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 \end{aligned}\right\} \text{----------(3)}$$

Equations (2) and (3) are known as the normal equations for estimating a and b. All the quantities $\sum_{i=1}^{n} x_i, \sum_{i=1}^{n} x_i^2, \sum_{i=1}^{n} y_i$ and $\sum_{i=1}^{n} x_i y_i$ can be obtained from the given set of points (xi,yi); i=1,2,.....n and the equation (2) and (3) can be solved for a and b. With the values of a and b so obtained equation (1) is the line of best fit to the given set of points (xi,yi); i=1,2,.....n.

Note: The equations of the line of best fit of y on x is obtained on eliminating a and b in (1) and (3) and can be expressed in the determinant form as follows

$$\begin{vmatrix} y & x & 1 \\ \sum y_i & \sum x_i & n \\ \sum x_i y_i & \sum x_i^2 & \sum x_i \end{vmatrix} = 0$$

Fitting of Second degree Parabola:

Let

$$y = a + bx + cx^2 \text{.....(1)}$$

Be the second degree parabola of best fit to set of 'n' points (xi,yi), i=1,2,.....n. Using the principle of least squares we have to determine a,b and c. So that

$$E=\sum_{i=1}^{n}\left[y_i-a-bx_i-cx_i^{\,2}\right]^2$$

Is minimum. Equating to zero the partial derivatives of E with respect to a,b, and c separately, we get the normal equation for estimating a,b and c as

$$\left.\begin{aligned}\frac{\partial E}{\partial a}=0\Rightarrow -2\sum_{i=1}^{n}\left[y_i-a-bx_i-cx_i^{\,2}\right]\\ \frac{\partial E}{\partial b}=0\Rightarrow -2\sum_{i=1}^{n}x_i\left[y_i-a-bx_i-cx_i^{\,2}\right]\\ \frac{\partial E}{\partial c}=0\Rightarrow -2\sum_{i=1}^{n}x_i^{\,2}\left[y_i-a-bx_i-cx_i^{\,2}\right]\end{aligned}\right\}\text{----------(2)}$$

$$\left.\begin{aligned}\Rightarrow \sum_i y_i=na+b\sum_i x_i+c\sum_i x_i^{\,2}\\ \sum_i x_iy_i=a\sum_i x_i+b\sum_i x_i^{\,2}+c\sum_i x_i^{\,3}\\ \sum_i x_i^{\,2}y_i=a\sum_n x_i^{\,2}+b\sum_n x_i^{\,3}+c\sum_n x_i^{\,4}\end{aligned}\right\}\text{----------(3)}$$

Summation taken over i from 1 to n. For given set of pionts (xi,yi); i=1,2,.....n, Equations(3) can be solved for a,b and c. And with these values of a,b and c equation (1) is the parabola of best fit.

Note: Eliminating a,b and c in (1) and (3) the parabola of best fit of y on x is given by

$$\begin{vmatrix} y & x^2 & x & 1\\ \sum y_i & \sum x_i^{\,2} & \sum x_i & n\\ \sum x_iy_i & \sum x_i^{\,3} & \sum x_i^{\,2} & \sum x_i\\ \sum x_i^{\,2}y_i^{\,2} & \sum x_i^{\,4} & \sum x_i^{\,3} & \sum x_i^{\,2}\end{vmatrix}=0\text{----------(4)}$$

Fitting of a power curve $y=ax^b$;

Fitting of a power curve of the forms

$$y=ax^b\text{----------(1)}$$

to a set of n pionts. Taking logarithms on both sides

we get

$$\log y=\log a+b\log x$$
$$\Rightarrow U=A+bV$$

where $U=\log y, A=\log a$ and $V=\log x$
$B=\log b$

this is a linear equation in V and U . Normal equation for estimating A and B are

$$\left.\begin{aligned} &\sum U = nA + b\sum V \\ &and \\ &\sum UV = A\sum V + b\sum V^2 \end{aligned}\right\} \text{----------(2)}$$

The equation of (2) can be solved for A and b. consequently we get a = antilogA . with the values of a and b so obtained equation (1) is the curve of best fit to the set of n points.

Exponential curves (i) $y = ab^x$ (ii) $y = ae^{bx}$

Fitting of exponential curves of the form to a set of n points

$$\text{(i)} \quad y = ab^x \text{.....(1)}$$

Taking logarithms on both sides of equation (1) we get

$$\log y = \log a + \log b$$
$$\Rightarrow U = A + BX$$

where $U = \log y, A = \log a$ and $B = \log b$ This are linear equation in X and U, the normal equations for estimating A and B are

$$\left.\begin{aligned} &\sum U = nA + B\sum X \\ &and \\ &\sum XU = A\sum X + B\sum X^2 \end{aligned}\right\} \text{----------(2)}$$

Solving these equations for A and B , we finally get

$a = anti\log A$ and $b = anti\log B$

with these values of a and b(1) is the curve of best fit to the given set of n points

(ii) Fitting of exponential curve of the form y=aebx to a set of 'n' points

$$y = ab^x \text{.....(1)}$$

Taking logarithm on both sides , we get

$$\left.\begin{aligned} &\log y = \log a + bx\log e \\ &\Rightarrow \log a + (b\log e)x \end{aligned}\right\} U = A + BX$$

where u=logy, A=log a and b = loge

this is linear equation in X and U

thus the normal equations are

$$\left.\begin{aligned} &\sum U = nA + B\sum X \\ &and \\ &\sum XU = A\sum X + B\sum X^2 \end{aligned}\right\} \text{----------(2)}$$

From these we find A and B and consequently

$a = anti\log(A)$ and $b = \dfrac{B}{\log e}$

Fitting of multiple regression model:

In multiple regression , we deal with data consisting of $n(r+1)-$ tuples $[x_{1i}, x_{2i}, ..., x_{ri}, y_i]$ where the x's are again assumed to be known without error while the y's are values of random variables . Data of this kind arise , for example , in studies designed to determine the effect of various climatic conditions on a metal's resistances to corrossion ;the effect of kiln temperature, humidity and iron content on the strength of a ceramic coating, on the effect of factory production , consumption level and stacks in storage on the price of a product.

As in the case of one independent variable, we shall first consider the problem where the regression equation is linear, namely, where for any given set of values $x_1, x_2, ..., x_r$ of the r independent variables, the mean of the distribution of y is given by

$$a_0 + a_1x_1 + a_2x_2 + ... + a_rx_r .....$$

For two independent variables, the problem of fitting a plane to a set of n points with coordinates (xi1,xi2,yi) . Applying the method of least squares to obtain estimates of the coefficients a0,a1 and a2 we minimise

$$y = a_0 + a_1x_1 + a_2x_2 + ... + a_rx_r ..... \text{----------(1)}$$

Let the residual be E_i

i.e; $E_i = [y - (a_0 + a_1x_{1i} + a_2x_{2_i} + ... + a_rx_{ri} .....)]$ ----------(2)

Suppose

$$S = \sum_{i=1}^{n} E_i^{\,2} = \sum_{i=1}^{n} [y - (a_0 + a_1x_{1i} + a_2x_{2_i} + ... + a_rx_{ri} .....)] \text{----------(3)}$$

The first order pratial derivatives of s with respect to a0,a1,a2,.....,ar should vanish.

$$\frac{\partial S}{\partial a_0} = 0, \frac{\partial S}{\partial a_1} = 0, \frac{\partial S}{\partial a_2} = 0, \frac{\partial S}{\partial a_3} = 0, ... \frac{\partial S}{\partial a_r} = 0 \text{ ----------(4)}$$

$$\left.\begin{aligned}
&-2\sum_{i=1}^{n}[y-(a_0+a_1x_{1i}+a_2x_{2_i}+.....+a_rx_{ri}.....)]=0\\
&-2\sum_{i=1}^{n}x_{1i}[y-(a_0+a_1x_{1i}+a_2x_{2_i}+.....+a_rx_{ri}.....)]=0\\
&-2\sum_{i=1}^{n}x_{2i}[y-(a_0+a_1x_{1i}+a_2x_{2_i}+.....+a_rx_{ri}.....)]=0\\
&.\\
&.\\
&.\\
&-2\sum_{i=1}^{n}x_{ri}[y-(a_0+a_1x_{1i}+a_2x_{2_i}+.....+a_rx_{ri}.....)]=0
\end{aligned}\right\}\text{----------(5)}$$

i.e,

$$\left.\begin{aligned}
&\sum_{i=1}^{n}y_i=na_0+a_1\sum_{i=1}^{n}x_{1i}+a_2\sum_{i=2}^{n}x_{2i}+.....+a_r\sum_{i=1}^{n}x_{ri}\\
&\sum_{i=1}^{n}x_{1i}y_i=a_0\sum_{i=1}^{n}x_{1i}+a_1\sum_{i=2}^{n}x_{i1}{}^2+a_2\sum_{i=2}^{n}x_{i1}x_{2i}+.....+a_r\sum_{i=1}^{n}x_{1i}x_{ri}\\
&\sum_{i=1}^{n}x_{2i}y_i=a_0\sum_{i=1}^{n}x_{2i}+a_1\sum_{i=2}^{n}x_{i1}x_{2i}+a_2\sum_{i=2}^{n}x_{2i}{}^2+.....+a_r\sum_{i=1}^{n}x_{2i}x_{ri}\\
&.\\
&.\\
&\sum_{i=1}^{n}x_{ri}y_i=a_0\sum_{i=1}^{n}x_{2i}x_{ri}+a_1\sum_{i=2}^{n}x_{ri}x_{1i}+.....+a_r\sum_{i=1}^{n}x_{ri}{}^2
\end{aligned}\right\}\text{----------(6)}$$

The set of equation (6) are (r+1) normal equation to fit

$$y=a_0+a_1x_1+a_2x_2+.....+a_rx_r$$

Suppose $y=a_0+a_1x_1+a_2x_2$, x1 and x2 are two independent variables the normal equations are

$$\left.\begin{aligned}
&\sum_{i=1}^{n}y_i=na_0+a_1\sum_{i=1}^{n}x_{1i}+a_2\sum_{i=2}^{n}x_{2i}\\
&\sum_{i=1}^{n}x_{1i}y_i=a_0\sum_{i=1}^{n}x_{1i}+a_1\sum_{i=2}^{n}x_{i1}{}^2+a_2\sum_{i=2}^{n}x_{i1}x_{2i}\\
&\sum_{i=1}^{n}x_{2i}y_i=a_0\sum_{i=1}^{n}x_{2i}+a_1\sum_{i=2}^{n}x_{i1}x_{2i}+a_2\sum_{i=2}^{n}x_{2i}{}^2
\end{aligned}\right\}\text{----------(7)}$$

19.3 MULTIPLE LINEAR REGRESSION - A MATRIX APPROACH TO LEAST SQUARES:

The model that we are the using in multiple linear regression lends itself uniquely to a unified treatment in matrix notation . In order to express the normal equations in matrix notation, Lets us define the following three matrices

$$X=\begin{bmatrix}1 & x_{11} & x_{12}\\ 1 & x_{21} & x_{22}\\ \vdots & \vdots & \vdots\\ 1 & x_{n1} & x_{n2}\end{bmatrix}, Y=\begin{bmatrix}y_1\\ y_2\\ \vdots\\ y_n\end{bmatrix} and B=\begin{bmatrix}a_0\\ a_1\\ \vdots\\ a_n\end{bmatrix}$$

The first one X, is an nx(2+1) matrix consisting essentially of the given values of x's , with column of 1's appended to accomodate the constant term . y is an nx1 matrix (or column vector) consisting of observes values the response variables and b is the (2+1)x1 matrix (or column vector) consisting of the least square estimates of the regression coefficients. The least squares estimates of the multiple regression coefficients are given by

$$B=(X'X)^{-1}X'Y$$

Where X' is the transpose of X and (X'X)-1 is the inverse of X'X. To verify this relation, we first determine X'X, X'XB and X'Y.

$$X'X=\begin{bmatrix}n & \sum X_{1i} & \sum X_{2i}\\ \sum X_{1i} & \sum X_{1i}^{\,2} & \sum X_{1i}X_{2i}\\ \sum X_{2i} & \sum X_{1i}X_{2i} & \sum X_{2i}^{\,2}\end{bmatrix}$$

$$X'XB=\begin{bmatrix}n & \sum X_{1i} & \sum X_{2i}\\ \sum X_{1i} & \sum X_{1i}^{\,2} & \sum X_{1i}X_{2i}\\ \sum X_{2i} & \sum X_{1i}X_{2i} & \sum X_{2i}^{\,2}\end{bmatrix}\begin{bmatrix}a_0\\ a_1\\ a_2\end{bmatrix}$$

$$=\begin{bmatrix}a_0+a_1X_{1i}+a_2X_{2i}\\ a_0\sum X_{1i}+a_1\sum X_{1i}^{\,2}+a_2\sum X_{1i}X_{2i}\\ a_0\sum X_{2i}+a_1\sum X_{2i}X_{1i}+a_2\sum X_{2i}^{\,2}\end{bmatrix}$$

$$X'Y=\begin{bmatrix}\sum Y_i\\ \sum X_{1i}Y_i\\ \sum X_{2i}Y_i\end{bmatrix}$$

Identifying the elements of X'XB as the expression on right hand side of the normal equation and those of X'Y as the expression on the left hand side, we can write

X'XB=X'Y

Multiplying on the both sodes of above equation by $(X'X)^{-1}$, we get

$$(X'X)^{-1}X'XB = (X'X)^{-1}X'Y$$

and finally we get,

$$B = (X'X)^{-1}X'Y$$

Since $(X'X)^{-1}X'X$ equals the (2+1)x(2+1) identity matrix I, amd by definition IB=B.

we have assumed here that X'X is non-singular , so that its inverse exists.

19.4 WORKED OUT EXAMPLES:

Example 1:

Find the parabola of the form $y = a + bx + cx^2$ which fits most closely with the observations

x: -3 -2 -1 0 1 2 3

y: 4.63 2.11 0.67 0.09 0.63 2.15 4.58

Solution: Normal Equation are

$$\sum_{i=1}^{n} Y_i = an + b\sum_{i=1}^{n} X_i + c\sum_{i=1}^{n} X_i^2$$

$$\sum_{i=1}^{n} X_iY_i = a\sum_{i=1}^{n} X_i + b\sum_{i=1}^{n} X_i^2 + c\sum_{i=1}^{n} X_i^3$$

$$\sum_{i=1}^{n} X_i^2Y_i = a\sum_{i=1}^{n} X_i^2 + b\sum_{i=1}^{n} X_i^3 + c\sum_{i=1}^{n} X_i^4$$

X_i	Y_i	X_iY_i	X_i^2	X_i^3	X_i^4	$X_i^2Y_i$
-3	4.63	-13.89	9	-27	81	41.67
-2	2.11	-4.22	4	-8	16	8.44
-1	0.67	-0.67	1	-1	1	0.67
0	0.09	0	0	0	0	0
1	0.63	0.63	1	1	1	0.63
2	2.15	4.30	4	8	16	8.60
3	4.58	13.74	9	27	81	41.22
0	14.86	-0.11	28	0	196	101.23

14.86=7a+28c.....(1)

-0.11=28b.....(2)

$\therefore b = \frac{-0.11}{28} = -0.004$

101.23=28a+196c.....(3)

equation(1) x4

==>59.44=28a+112c.....(4)

Solving eq(3) & (4) we get

41.79=84c

$\therefore c = 0.5$

a=14.86-28c/7==>a+14.86-28(0.5)/7==> $\therefore a = 0.12$

$\therefore$ y=0.12-0.004 x +0.5 x^2 is the required parabola

EXAMPLE 2:

Fit an equation of the form y=abx to the following data

X:	2	3	4	5	6
Y:	144	172.8	207.4	248.8	298.5

Solution: Equation of the form $y = ab^x$ .....(1)

Taking logarithms on both sides of eq(1) we get

$$\log y = \log a + x \log b$$

then $U = A + BX$ .....(2)

where $U = \log y$, $A = \log a$, $B = \log b$

The normal equation are

$\sum U = nA + B\sum X$ and

$\sum XU = A\sum X + B\sum X^2$

X	Y	$U = \log y$	X^2	XY
2	144	2.1584	4	4.3168
3	172.8	2.2375	9	6.7125
4	207.4	2.3168	16	9.2672
5	248.8	2.3959	25	11.9795
6	298.5	2.4749	36	14.8494
20		11.5835	90	47.1254

11.5835=5A+20B.....(1)

47.1254=20A+90B.....(2)

eq(1)x4 46.3340=20A+80B

47.1254=20A+90B

0.7914=10B==> $\boxed{B = 0.07914}$

A=11.5835-20B/5 ==> $\boxed{A = 2}$

Since $A = \log a ==> a = anti\log A$ =Antilog2=100

$B = \log b ==> b = anti\log B$ =antilog(0.07914)=1.2

Since $y = ab^x$

Hence $\boxed{Y = 100(1.2)^X}$

EXAMPLE 3:

Find the curve of best fit of the type y=aebx to the following data by the method of least squares.

X:	1	5	7	9	12
Y:	10	15	12	15	21

Solution: The curve is $y = a.e^{bx}$ .....(1)

taking logarithms on both sides

$\log Y = \log a + bX \log e$

i.e; $U = A + BX$

where $U = \log y$, $A = \log a$, $B = b\log e$

normal equations are

$\sum U_i = nA + B\sum X$

$\sum X_i U_i = A\sum X_i + B\sum X_i^2$

Xi	Yi	$U = \log y$	X_i^2	X_iY_i
1	10	1.0000	1	1
5	15	1.1761	25	5.8805
7	12	1.0792	49	7.5544
9	15	1.1761	81	10.5849
12	21	1.3222	144	15.8664
34		5.7536	300	40.8862

n=5, we have by using above normal equations

5.7536 = 5A+34B.....(2)

40.8862 = 3A+300B.....(3)

eq(2)x34 ==>195.6224 = 170A+1156B.....(4)

eq(2)x5==> 204.4310 = 170A+1500B.....(5)

Solving(4)&(5) equations

8.8086 = 344B

B = 8.8086/344 ==> 0.0256

A=5.7536-34(0.0256)/5 ==>0.97

a=9.4754, b=0.059

∴ y=9.4754 e0.059

EXAMPLE 4:

Fit the model y=axb to the following data

X:	1	2	3	4	5	6
Y:	2.98	4.26	5.21	6.10	6.80	7.50

Solution: The curve is $y = ax^b$ .....(1)

Taking logarithms on both sides we get

$\log y = \log a + b \log x$

$U = A + bV$

where U=logy, A=loga, V=logX

∴ normal equations are

$\sum U_i = nA + b\sum V_i$

$\sum U_i V_i = A\sum V_i + b\sum V_i^2$

X	Y	V=logX	U=logY	UV	V2
1	2.98	0	0.4742	0	0
2	4.26	0.3010	0.6294	0.1894	0.0906
3	5.21	0.4771	o.7168	0.3420	0.2276
4	6.10	0.6021	0.7853	0.4728	0.3625
5	6.80	0.6990	0.8325	0.5819	0.4886
6	7.50	0.7782	0.8751	0.6810	0.6056
		2.8574	4.3133	2.2671	1.7749

∴ The normal equation are

4.3133=6A+2.8574b.....(2)

2.2671=2.8574A+1.7749b.....(3)

By Solving above two equations we get

B=0.5142, A=0.4740 ==>loga=0.4740

==> a=Antilog(0.4740)

==> a=2.978

$\therefore y = ax^b$ =2.978x0.5142

EXAMPLE 5:

Find the least squares regression equation of X1 on X2 and X3 from the following data.

X1:	3	5	6	8	12	14
X2:	16	10	7	4	3	2
X3:	90	72	54	42	30	12

Solution: Let $x_1 = a_0 + a_1x_2 + a_2x_3$ ……..(1)

Changing the origin $u = x_2 - 7$ and $v = x_3 - 50$

Let $x_1 = a + bu + cv$ ……..(2)

Then normal equations are

$$\sum_{i=1}^{n} x_{1i} = na + b\sum_{i=1}^{n} u_i + c\sum_{i=1}^{n} v_i$$

$$\sum_{i=1}^{n} x_{1i}u_i = a\sum_{i=1}^{n} u_i + b\sum_{i=1}^{n} u_i^2 + c\sum_{i=1}^{n} u_iv_i$$

$$\sum_{i=1}^{n} x_{1i}u_i = a\sum_{i=1}^{n} u_i + b\sum_{i=1}^{n} u_iv_i + c\sum_{i=1}^{n} v_i^2$$

X1	X2	X3	Ui	Vi	X1iUi	X1iVi	UiVi	Ui2	Vi2
3	16	90	9	40	27	120	360	81	1600
5	10	72	3	22	15	110	66	9	484
6	7	54	0	4	0	24	0	0	16
8	4	42	-3	-8	-24	-64	24	9	64
12	3	30	-4	-20	-48	-240	80	16	400
14	2	12	-5	-38	-70	-532	190	25	1444
48			0	0	-100	-582	720	140	4008

Here n=6,

48=6a+0+0 ==> a=8…….(1)

-100=140b+720c…….(2)

-582=720b+4008c……(3)

(2) x36 ==> -3600=5040b+2920c

==>-4074=5040b+28056c

474=-2136c

$\therefore$ c=-0.22

B=100-720(-0.22)/140=0.417

Therefore y=8+0.417U-0.22V

==> 8+0.417(X2-7) - 0.22(X3-50)

==> 8+0.417X2-2.919-0.22X2+11

$y = 16.1 + 0.417x_2 - 0.22x_3$

EXAMPLE 6:

The following data show the number of bedrooms the number of baths , and the prices at which a random sample of eight one family houses sold recently in a certain large housing development.

Number of bedrooms x_1	Number of baths x_2	Price(dollars y)
3	2	78800
2	1	74300
4	3	83800
2	1	74200
3	2	79700
2	2	74900
5	3	88400
4	2	82900

Use the method of least squares to find a linear equation which will enable as to predict the average the average sales price of a one family house in the given housing development in terms of the number of bedrooms and the number of baths.

Solution : Let $y = a_0 + a_1x_1 + a_2x_2$ .....(1)

the normal equations are

$$\sum_{i=1}^{n} y_i = na_0 + a_1\sum_{i=1}^{n} x_{1i} + a_2\sum_{i=1}^{n} x_{2i} \text{ .....(2)}$$

$$\sum_{i=1}^{n} x_{1i}y_i = a_0\sum_{i=1}^{n} x_{1i} + a_1\sum_{i=1}^{n} x_{1i}^{\,2} + a_2\sum_{i=1}^{n} x_{1i}x_{2i} \text{ ……..(3)}$$

$$\sum_{i=1}^{n} x_{2i}y_i = a_0\sum_{i=1}^{n} x_{2i} + a_1\sum_{i=1}^{n} x_{1i}x_{2i} + a_2\sum_{i=1}^{n} x_{2i}^{\ 2} \text{ .....(4)}$$

where n = 8 , $\sum_{i=1}^{n} x_{1i} = 25$, $\sum_{i=1}^{n} x_{2i} = 16$, $\sum_{i=1}^{n} y_i$ =637000

$\sum_{i=1}^{n} x_{2i}^{\ 2} = 87$, $\sum_{i=1}^{n} x_{1i}x_{2i} = 55$, $\sum_{i=1}^{n} x_{2i}^{\ 2} = 36$, sigma $\sum_{i=1}^{n} x_{1i}x_{2i}$ =2031000

and $\sum_{i=1}^{n} x_{2i}y_i$ =129770 and we get from (2) (3) (4)

637000=8 a_0 +25 a_1 +61 a_2 .....(5)

2031100=25 a_0 +87 a_1 +55 a_2 .....(6)

1297700=16 a_0 +55 a_1 +36 a_2 .....(7)

Solving (5) (6) (7) equations by the method of elimination we get the coefficient values a_0 =65191.7 a_1 =4133.3 and a_2 =758.3 then the least squares equation (1) becomes

y=65192+4133 x_1 +758 x_2 .....(8)

Which tells us that in the given housing development and at the time the study was made each extra

$\sum_{i=1}^{n} x_{1i}x_{2i}$ =500 , $\sum_{i=1}^{n} x_{2i}^{\ 2}$ =3000 and n=16

Are substituted in to the expression of X'X given by

$$X'X = \begin{bmatrix} n & \sum x_{1i} & \sum x_{2i} \\ \sum x_{1i} & \sum x_{1i}^{\ 2} & \sum x_{1i}x_{2i} \\ \sum x_{2i} & \sum x_{1i}x_{2i} & \sum x_{2i}^{\ 2} \end{bmatrix} = \begin{bmatrix} 16 & 40 & 200 \\ 40 & 120 & 500 \\ 200 & 500 & 3000 \end{bmatrix}$$

Then the inverse of their matrix can be obtained by any one of a number of technique and using the one based on co factors we find that

$$(XX)^{-1} = \frac{1}{1,60,000}\begin{bmatrix} 1,10,000 & -2000 & -4000 \\ -20,000 & 8000 & 0 \\ -4000 & 0 & 320 \end{bmatrix}$$

Where 160000 is the value of $(X'X)$, the determinant of $X'X$. Substituting $\sum y_i$ =723 $\sum x_{1i}y_i$ =1963 and $\sum x_{2i}y_i$ =8210 in the expression of $X'Y$ we then get

$$X'Y = \begin{bmatrix} \sum y_i \\ \sum x_{1i} y_i \\ \sum x_{2i} y_i \end{bmatrix} = \begin{bmatrix} 723 \\ 1913 \\ 8210 \end{bmatrix}$$

And finally

$$B = (X'X)^{-1} XY = \frac{1}{16000} \begin{bmatrix} 1,10,000 & -20,000 & -4000 \\ -20,000 & 81000 & 0 \\ -4000 & 0 & 320 \end{bmatrix}$$

$$= \frac{1}{16,000} \begin{bmatrix} 74,30,000 \\ 12,44,000 \\ -2,64,800 \end{bmatrix}$$

$$= \begin{bmatrix} 46.4375 \\ 7.7750 \\ -1.6550 \end{bmatrix}$$

Bedroom adds on the average 4133 dollars and each bath 758 dollars to the sales price of a house. Hence based on the result obtained above we predict the sales price of a three bedroom house with two baths in the large housing development.

Substituting $x_1 = 3$ and $x_2 = 2$ into the equation(8) obtained above , we get

$\hat{y} = 65,192 + 4,133(3) + 758(2)$

$\hat{y} = 5,79,107$

EAMPLE 7:

The following are data on the number of twists required to break a certain kind of forged alloy bar and the percentages of two alloying elements present in the metal with the following data.

$\sum_{i=1}^{n} x_{1i} = 40$ $\sum_{i=1}^{n} x_{2i} = 200$ $\sum_{i=1}^{n} x_{1i}^2 = 120$ $\sum_{i=1}^{n} x_{1i} x_{2i} = 500$

$\sum_{i=1}^{n} x_{2i}^2 = 3000$ $\sum_{i=1}^{n} y_i = 723$ $\sum_{i=1}^{n} x_{1i} y_i = 1963$

$\sum_{i=1}^{n} x_{2i} y_i = 8210$ with normal equations

$16 a_0 + 40 a_1 + 200 a_2 = 723$

$40 a_0 + 120 a_1 + 500 a_2 = 1963$

$200\,a_0+500\,a_1+3000\,a_2=8210$

Use the matrix expressions to determine the least squares estimates of the multiple regression coefficients.

Solution: Since we are given $\sum_{i=1}^{n} x_{1i}=40$ $\sum_{i=1}^{n} x_{2i}=200$ $\sum_{i=1}^{n} x_{1i}^{\,2}=120$

$$B=(X'X)^{-1}X'Y=\begin{bmatrix}46.4375\\7.7750\\-1.6550\end{bmatrix}$$

EXAMPLE 8:

The following data show the number of bedrooms and baths and the prices at which a random sample of eight one family houses sold recently in a certain large housing development determine the value of variance^ for the data.

Number of Bedrooms x_1	Number of Baths x_2	prices(Dollars y)
3	2	78800
2	1	74300
4	3	83800
2	1	74200
3	2	79700
2	2	74900
5	3	88400
4	2	82900

Using matrix expressions to determine the least squares estimates of the multiple regression coefficients.

Solution: the quantities we need for substitution into the expression for X'X is n=8

$$\sum_{i=1}^{n} x_{1i}=25\quad \sum_{i=1}^{n} x_{2i}=16\quad \sum_{i=1}^{n} x_{1i}^{\,2}=\quad 87\quad \sum_{i=1}^{n} x_{1i}x_{2i}=55\quad \sum_{i=1}^{n} x_{2i}^{\,2}=36$$

we get

$$(X'X)=\begin{bmatrix}8 & 25 & 16\\25 & 87 & 55\\16 & 55 & 36\end{bmatrix}$$

Then the inverse of the matrix can be obtained by any of a number of different techniques using the one based on cofactors we find that

$$(X'X) = \begin{bmatrix} 107 & -20 & -17 \\ -20 & 32 & -40 \\ -17 & -40 & 71 \end{bmatrix} \frac{1}{84}$$

Where 84 is the value of $|\ |X'X|$, the determinmant of $X'X$. Substituting sigma y = 637000, $\sum_{i=1}^{n} x_{1i}y_i = 2031100$ and $\sum_{i=1}^{n} x_{2i}y_i = 1297700$ into the expression for $X'Y$ we then get

$$X'Y = \begin{bmatrix} 6,37,000 \\ 2,00,31,100 \\ 12,97,700 \end{bmatrix}$$

And finally

$$(X'X)^{-1}X'Y = \frac{1}{84}\begin{bmatrix} 107 & -20 & -17 \\ -20 & 32 & -40 \\ -17 & -40 & 71 \end{bmatrix}\begin{bmatrix} 6,37,000 \\ 20,31,100 \\ 12,97,700 \end{bmatrix}$$

$$= \frac{1}{84}\begin{bmatrix} 54,76,100 \\ 3,47,200 \\ 63,700 \end{bmatrix}$$

$$= \begin{bmatrix} 65,191.7 \\ 4,133.3 \\ 758.3 \end{bmatrix}$$

Where the $a_0 = 65191.7$, $a_1 = 4133.3$, $a_2 = 758.3$

First calculate $Y'Y$, which is simply $\sum_{i=1}^{n} y_i^2$, so we get

$$Y'Y = 78800^2 + 74300^2 + \ldots\ldots + 82900^2$$

$$= 50,907,080,000$$

The value of $B' = \frac{1}{84}[54,76,100 \quad 3,47,200 \quad 63,700]$

$$X'Y = \begin{bmatrix} 6,37,000 \\ 20,31,100 \\ 12,97,700 \end{bmatrix}$$

$$\therefore B'X'Y = \frac{1}{84}[54,76,100 \quad 3,47,200 \quad 63,700]\begin{bmatrix} 6,37,000 \\ 20,31,100 \\ 12,97,700 \end{bmatrix}$$

$$=50,906,394,166$$

And follows that $\hat{\sigma} = \sqrt{\frac{50906394166 - 50906394166}{8}}$

$$\hat{\sigma} = 292.8$$

19.5 EXERCISE

1. Fit the curve $y=a.e^{bx}$ to the following data

X : 0.0 0.5 1.0 1.5 2.0 2.5

Y: 0.10 0.45 2.15 9.15 40.35 180.75

[ANS: $y=0.1019e^{2.9963x}$]

2. Fit $y=a.b^x$ by the method of least squares to the following data

X: 0 1 2 3 4 5 6 7

Y: 10 21 35 59 92 200 400 610

[ANS: $Y=10.499(1.7959)^X$]

3. Fit a parabola for the data

X: 1 2 3 4 5

Y:1090 1220 1390 1625 1915

[ANS: $Y=1024+40.5X+27.5X^2$]

4. Fit a straight line y=a+bx and also a parabola to the following set of observations the sum of squares of residuals in each case and test which curve is more suitable to the data

X: 0 1 2 3 4

Y: 1 5 10 22 38

[ANS: Y=9.1X-3, Y=1.42+0.26X is the parabola of best fit.+$2.21x^2$]

5. The following sample data where collected to determine the relationship between two processing variables and the current gain of a certain kind of transistor fit the least squares regression equation of y on x1 and x2 and use the matrix rotation also

X1: 1.5 2.5 0.5 1.2 2.6 0.3 2.4 2.0 0.7 1.6

X2:66 87 69 141 93 105 111 78 66 123

Y:5.3 7.8 7.4 9.8 10.8 9.1 8.1 7.2 6.5 12.6

[ANS: y=2.3+0.23x1+0.06x2]

6. The following are sample data provided by a moving company on the wights of six shipments, the distances they were moved and the damaged that was incurred.

Weights(1000lb)	Distance(1000 miles)	Damage(Dollars)
X1	x2	y
4.0	1.5	160
3.0	2.2	112
1.6	1.0	69
1.2	2.0	90
3.4	0.8	123
4.8	1.6	186

(a) Assuming that the regression is linear, estimate a_0, a_1 and a_2

(b) Use the resolts of (a) to estimate the damage when a shipment weighing 2,400 pounds is moved 1,200 miles

7. Use the matrix relations to fit a straight line to the data

X: 0 1 2 3 4

Y:8 9 4 3 1

8. The following are data on the number of twists required to break a certain kind of forged alloy bar and the percentages of two alloying elements present in the metal

No.of twists	percent of elements	percent of elements
y	x_1	x_2
41	1	5
49	2	5
69	3	5
65	4	5
40	1	10
50	2	10

58	3	10
57	4	10
31	1	15
36	2	15
44	3	15
58	4	15
19	1	20
32	2	20
33	3	20
42	4	20

Use the matrix expressions to determine the least squares estimates of the multiple regression coefficients.

ANS: $B = \begin{bmatrix} 46.4375 \\ 7.7750 \\ -1.6550 \end{bmatrix}$

9. Define multiple regression and discuss a procedure to fit the curve

$y = a_0 + a_1 x_1 + a_2 x_{2i}^{\ 2}$

19.6 SUMMARY

In this lesson an attempt is made to explain the concepts of multiple regression models method of least squares procedure associated with them along with theory and practical. A number of examples are worked out and a good number of exercises are also given.

19.7 TECHNICAL TERMS.

- Multiple linear regression
- Least squares procedures
- Matrix approach to least square procedures.

19.8 SELF ASSESSMENT QUESTIONS

SHORT:

1. What is the purpose of the least squares procedure in multiple linear regression?
2. Define multiple linear regression and explain its basic model equation.

3. How does the matrix approach simplify the computation of regression coefficients?
4. What are the key assumptions behind the least squares method in multiple regression?
5. Briefly describe how worked-out examples help in understanding multiple linear regression models.

ESSAY:

1. Explain the concept of multiple linear regression and describe the least squares procedure for model fitting.
2. Discuss the advantages of using a matrix approach to solve multiple linear regression problems.
3. Describe the step-by-step process of fitting a multiple linear regression model using the least squares method.
4. Evaluate the importance of verifying the underlying assumptions in multiple linear regression analysis.
5. Analyze a real-world dataset using multiple linear regression: Formulate the model, apply the matrix approach for coefficient estimation, and interpret the results.

19.9 FURTHER READING

(1) "Introduction to probability and statistics" by J. Susan Milton and J.C. Arnold, 4th edition, TMH (2007)

(2) "Mathematical Statistics" by R.K. Goyal, Krishna Prakashan Media (P) Ltd, Meerut.

(3) "Fundamentals of Mathematical Statistics" by S.C. Gupta and V.K. Kapoor, S.Chand & Sons, New Delhi

Dr. B. Sri Ram

LESSON-20

MULTIPLE LINGAR REGRESSION MODELS-INTERVAL ESTIMATION

OBJECTIVE:

This lesson is prepared in such a way that after studying the material the student is expected to have a thorough comprehension of the concept "Multiple Linear Regression Models-Interval Estimation" the breath of statistical inference and analysis. The student would be equipped with theoretical as well as practical aspects of concepts.

STRUCTURE OF THE LESSON:

20.1 Introduction

20.2 Multiple Linear Regression Models - Interval Estimation

20.3 Worked out Examples

20.4 Exercise

20.5 Summary

20.6 Technical Terms

20.7 Self assessment questions

20.8 Further Reading

20.1 INTRODUCTION:

A model for an experiment must have parameters, the unknown of the experiment. The purpose of performing an experiment is always to get information about the unknown parameters. In finding out an unknown population parameter, a judgement or statement is usually made which is only an estimate. For example sample mean is an estimator of population mean because, sample mean is. a method of determining the population mean consider a random sample $x_1,x_2...x_n$ from a population with population density function $f(x,\theta)$, where 'θ' is the unknown parameter. In order to find the estimate in terms of sample values, we have to construct a number of statistical functions of $x_1,x_2,\ldots x_n$ which is said to be estimate of 'θ'. For a parameter, there can be a number of estimators. The main part of

estimation problem lies in selecting out sample functions or estimators $\hat{\theta}_1$, $\hat{\theta}_{2....}$ $\hat{\theta}_n$ such that their distributions are concentrated as closely as possible round the time parameter values θ_1, $\theta_{2,....}$ θ_n respectively. The estimation can be done in either of the two ways (i) Point Estimation (ii) Interval Estimation.

A single-valued estimate or a point estimate does not, in general coincide with a true value of the parameter. It is preferred to obtain a range of values or an interval, thus the procedure of determining an interval (a,b) that will include a population parameter, says θ, with a certain probability (1-α) is known as internal estimation the main objective of many statistical study is to make predictions, preferably on the basis of mathematical equations for instance, an engineer may wish to predict the amount of oxide that will form on the surface of a metal baked in an oven for one hour at 200 degrees Celsius. Usually, such predictions require that a formula be found which relates the dependent variable whose value one wants to predict to one or more independent variables. Hence, a descriptive explanation for the concept of interval estimation in multiple linear regression was given. In this Lesson we discuss at length the interval estimation for multiple linear regression models with specific applications.

20.2 MULTIPLE LINEAR REGRESSION MODELS - INTERVAL ESTIMATION:

In multiple regression, the objective is to build a probabilistic model that relates a dependent variable y to more than one independent or predictor variable. Let k represent the number of predictor variables ($k \geq 2$) and denote these predictors by $x_1, x_2, ...x_n$. For example, in attempting to predict the selling price of a house, we might have k=3 with x_1= size(sq++), x_2= age (years) and x_3 = number of rooms. Then the general additive multiple regression model equation is

$$Y= \beta_0+\beta_1 x_1+ \beta_2 x_n+....+ \beta_n x_n+\varepsilon \qquad \rightarrow(1)$$

Where $E(\varepsilon) = 0$ $V(\varepsilon)= \sigma^2$. In addition, for purposes of testing hypothesis and calculating confidence intervals. or predictive intervals, it is assumed that ε is normally distributed.

Thus just as $\beta_0+\beta_1 x$ describes the mean y value as a function of x in simple linear regression, the true or population regression function $\beta_0+\beta_1 x_1+....+ \beta_k x_k$ gives the expected value of y as a function of $x_1...x_k$. The β_i's are the true or population regression coefficients.

The regression coefficient β_1 is interpreted as the expected change in y associated with a 1-unt increase in x_1 while x_2,....x_k are held fixed.

Constructing confidence intervals, and making predictions, one should first examine diagnostic plots to see whether the model needs modification or whether there are outliers in the data. the recommended plots are standardised residuals versus each independent variable, residuals versus $\bar{y}$, $\bar{y}$ versus y and a normal probability plot of the standardised residuals. Potential problems are suggested by the same patterns of particular importance is the identification of observations that have a large inference on the fit. Because each $\hat{\beta}$, is a linear function of the y_i , s the standard deviation of each $\hat{\beta}$, is the product of σ and a function of the x_{ij}s, so an estimate $S\hat{\beta}$ is obtained by substituting s for σ. The function of the x_{ij}s is quite complicated, but all standard regression computer packages compute and show the $S\hat{\beta}$iS. Inference concerning a single β , are based on the standardised variable.

$$T=\frac{\hat{\beta}i-\beta i}{S\,\hat{\beta}i}$$

Which has a t-distribution with n-(k+1) df

The point estimate of $\mu_y /x_1^*, x_2^*.... x_k^*$ the expected value of y when $x_1 = x_1^*,.... x_{k=} x_k^*$ is $\hat{\mu}_y/x_1^*.... x_k^*= \hat{\beta}_0 + \hat{\beta}_1 x_1^* + \beta_2 x_n +....+ \hat{\beta}_k x_k^*$.

The estimated standard deviation of the corresponding estimator is again an expression involving the sample x_{ij}s. Inferences about $\mu_y /x_1^*, ... x_k^*$ are based on standardizing its estimator to obtain a t-variable having n-(k+1) d.f.

Therefore, a 100 (1-α)% C.I for β, the coefficient of x, in the regression function is S_y

$$\hat{\beta}\pm t_{\alpha/2},\ n\text{-}k+1.\ S\,\hat{\beta}_i$$

A test for H_0:$\beta= \beta_0$ uses the t-statistic value $\bar{t} = \frac{(\hat{\beta}i-\hat{\beta}io)}{s\hat{\beta}i}$ based on n-(k + 1)d.f

The test is upper, lower, or two tailed according to whether H_0 contains the inequality >,< or ≠

A 100(1-α)% confidence interval for $\hat{\mu}_y/x_1^*.... x_k^*\pm t_{\alpha/2},n\text{-}k+1$[estimated S.D of $\hat{\mu}_y/x_1^*.... x_k^*$]

$$=\hat{y}\pm t_{\alpha/2},\ n\text{-}k+1.\ S_{\hat{y}}$$

Where $\hat{y}$ is the statistic $\hat{\beta}_0 + \hat{\beta}_1 x_1^* + \beta_2 x_n +....+ \hat{\beta}_k x_k^*$ and $\hat{y}$ is the calculated value of $\hat{y}$.

Also, a 100(1-α)%, predictive interval for a future y-value is

$\hat{\mu}_y/x_1^*.... x_k^*\pm t_{\alpha/2},\ n\text{-}k+1[\ S^2+$(estimated S.D of $\hat{\mu}_y/x_1^*.... x_k^*)^2]^{1/2}$

$= \hat{y}\pm t_{\alpha/2},\ n\text{-}k+1.\sqrt{S^2 + Sy^2}$

Simultaneous intervals for which the simultaneous confidence or prediction level is controlled. Therefore under the normality assumption.

$$\hat{\beta} \sim MN_{(}\beta,\sigma^2_{(X^1X)^{-1}})$$

$$cov_{(}\hat{\beta}_{)=}\ \sigma^2_{(X^1X)^{-1}=}\ \sigma^2 c$$

$var(\hat{\beta}_j)= \sigma^2 c_{jj}$

Standard Error: $S.E(\hat{\beta}_j)= \sqrt{\hat{\sigma}2\ cjj}$

$\beta_j \in \hat{\beta}_j \pm t_{\alpha/2,\ n-p} \sqrt{\hat{\sigma}2cjj}$

20.3 WORKED OUT EXAMPLES

Example 1:

Soil and sediment adsorption, the extent to which chemicals collect in a condensed form on the surface, is an important characteristic influencing the effectiveness of pesticides and various agricultural chemicals. The article "Adsorption of Phosphate, Arsenate, Methane arsonate and cacodylate by lake and stream sediments: comparisons with soils", gives the accompanying data on y = phospate adsorption index, x_1 = amount of extractable iron and x_2 = amount of extractable aluminium.

observation:	1	2	3	4	5	6	7	8	9	10	11	12	13
Extractable Iron=x_1:	61	175	111	124	130	173	169	169	160	244	257	333	199
Extractable Aluminium=x_2:	13	21	24	23	64	38	33	61	39	71	112	88	54
Adsorption Index:	4	18	14	18	26	26	21	30	28	36	65	62	40

Solution: The proposed model is $Y= \beta_0+\beta_1x_1+ \beta_2x_n+....+ \beta_nx_n+\varepsilon$

for parameter β_i :

β_0 estimate $\hat{\beta}_0$: -7.351, β_1 estimate $\hat{\beta}_1$: 0.11273, β_2 estimate $\hat{\beta}_2$: 0.34900 } Estimate of $\hat{\beta}_i$

$SD(\beta_0)=S\hat{\beta}_0=3.485$, $SD(\beta_1)=S\hat{\beta}_1=0.02969$, $SD(\beta_2)=S\hat{\beta}_2=0.07131$} Estimated SD $S\hat{\beta}_i$

$R^2=0.948$, adjusted $R^2=0.938$, $S=4.379$.

$\hat{\mu}_{y.160.39}=\hat{y}=-7.351+(.11273)(160)+(0.34900)(39)=24.30$

Estimated SD of $\hat{\mu}_y=S_{\hat{y}}=1.30$

A 100 (1-α)% C.I for β, the coefficient of x, in the regression function is S_y

$$\hat{\beta}_i \pm t_{\alpha/2,\ n-(k+1)}.\ S\hat{\beta}_i$$

The 99% CI for β,the change in expected adsorption associated with a 1-unit increase in extractable iron while. extractable aluminium is held fixed, requires $t_{0.05,13-(2+1)=t}$ $t_{0.05,10}$=3.169
The CI is
0.11273±(3.169)(0.02969)=0.11273±0.09409=(0.019,0.207)

99% CI for β_2 is 0.34900± (3.169) (0.07131)=0.34900±0.22598=(0.123,0.575)

The 95% CI for $\mu_{y,160,39}$ expected adsorption when extractable iron = 160 and extractable aluminium=39 is 24.30± (2.228) (1.30)= 24.30±2.90=(21.40,27.20)
The 95% Predictive interval for a future value of adsorption to be observed when x_1= 160 and x_2 = 39 is $24.30\pm(2.228)\{(4.379)^2+1.30)^2\}^{1/2}$

=24.30±10.18=(14.12,34.48).

Example 2:

Find 95% confidence interval for β_1, given $\hat{\beta}_1$= 1.61591 c_{11} = 0 0.0027438, $\hat{\sigma}^2$=10.6239, $t_{0.025,22}$=2.074 .

Solution: The Confidence interval for multiple regression is given by

$$\hat{\beta}_j \pm t_{\alpha/2},_{n-p} \sqrt{\hat{\sigma}^2 c_{jj}}$$

Or $\beta_1 \in [\ \hat{\beta}_j \pm t_{\alpha/2},_{n-p}\sqrt{\hat{\sigma}^2 c_{jj}}\ ,\ \hat{\beta}_j \pm t_{\alpha/2},_{n-p} \sqrt{\hat{\sigma}^2 c_{jj}}\ ]$ →(1)

Where $\hat{\beta}_1$=1.61591, c_{11} =0.0027438, $\hat{\sigma}^2$=10.6239
$t_{0.025,\ 22}$ =2.074.
using these values in equation (1) we get

$\beta_1 \in [\ \hat{\beta}_1 \pm t_{\alpha/2},_{n-p}\sqrt{\hat{\sigma}^2 c_{jj}}\ ,\ \hat{\beta}_1 \pm t_{\alpha/2},_{n-p} \sqrt{\hat{\sigma}^2 c_{jj}}\ ]$

$\Longrightarrow \beta_1 \in [\ \hat{\beta}_1 \pm t_{0.025,22}\sqrt{\hat{\sigma}^2 c_{11}}\ ,\ \hat{\beta}_1 \pm t_{0.025},_{22} \sqrt{\hat{\sigma}^2 c_{11}}\ ]$

$\Longrightarrow \beta_1 \in [\ 1.61591 - 2.074\sqrt{10.6239 \times 0.0027438}\ ,$

$1.61591 + 2.074 \times \sqrt{10.6239 \times 0.0027438}\ ]$

$\therefore \beta_1 \in [1.26181,1.97001]$ or $1.26181 < \beta_1 < 1.97001$

Which is the required confidence interval.

20.4 EXERCISE

1. A multiple regression analysis. was carried out to relate y = tensile strength of a synthetic-fiber specimen to the variables x_1 = percent cotton and x_2= drying time. The data set consisted of n= 12 observations. The estimated standard deviation of $\hat{\beta}_0 + \hat{\beta}_1 x_1 + \beta_2 x_2$ when x_1=18 and x_2=3.0 was 1.20 calculate a 95% predictive interval for tensile strength of a fabric specimen for which $x_1 = 18$ and $x_2 = 3.0$

20.5 SUMMARY

In this lesson an attempt is made to explain the concepts of Multiple regression models-interval estimation procedures associated with them along with both theory and practical. The respective examples pertaining to above said interval estimation are worked out and a good number of exercise problems are given.

20.6 Technical Terms

- Multiple Linear regression.
- Interval estimation
- Confidence Intervals.
- Predictive Intervals
- Level of Significance.
- Degrees of freedom
- Critical value.

20.7 Self assessment questions

SHORT:

1. What is interval estimation in the context of multiple linear regression models?
2. Define the concept of a confidence interval for regression coefficients.
3. How is the standard error used in constructing interval estimates in regression analysis?
4. What does a 95% confidence interval indicate for a regression parameter?
5. Briefly describe the steps involved in a worked-out example of interval estimation in multiple linear regression.

ESSAY:

1. Explain the process of interval estimation in multiple linear regression models, detailing how confidence intervals for regression coefficients are constructed.
2. Discuss the importance of interval estimation in regression analysis and how it contributes to assessing the reliability of estimated regression parameters.
3. Derive the formula for a confidence interval for a regression coefficient in a multiple linear regression model.
4. Analyze a worked-out example of interval estimation in a multiple linear regression model.
5. Critically evaluate the limitations of interval estimation in multiple linear regression, particularly in the presence of issues like multicollinearity and heteroscedasticity.

20.8 Further Reading

(1) "Introduction to probability and statistics" by J. Susan Milton and J.C. Arnold, 4th edition, TMH (2007)

(2) "Mathematical Statistics" by R.K. Goyal, Krishna Prakashan Media (P) Ltd, Meerut.

(3) "Fundamentals of Mathematical Statistics" by S.C. Gupta and V.K. Kapoor, S.Chand & Sons, New Delhi.

Dr. B. Sri Ram

STATISTICAL TABLES

Cumulative normal distribution

Critical values of the *t* distribution

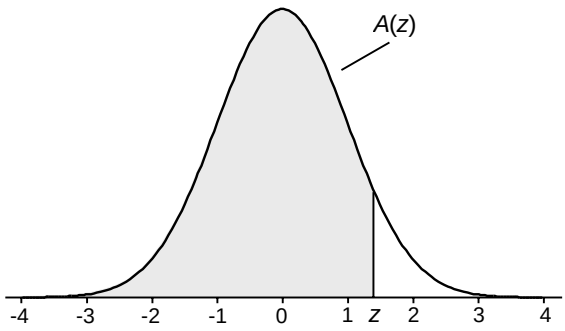
Critical values of the *F* distribution

Critical values of the chi-squared distribution

.

TABLE A.1

Cumulative Standardized Normal Distribution



$A(z)$ is the integral of the standardized normal distribution from $-\infty$ to z (in other words, the area under the curve to the left of z). It gives the probability of a normal random variable not being more than z standard deviations above its mean. Values of z of particular importance:

z	$A(z)$	
1.645	0.9500	Lower limit of right 5% tail
1.960	0.9750	Lower limit of right 2.5% tail
2.326	0.9900	Lower limit of right 1% tail
2.576	0.9950	Lower limit of right 0.5% tail
3.090	0.9990	Lower limit of right 0.1% tail
3.291	0.9995	Lower limit of right 0.05% tail

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998
3.5	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998
3.6	0.9998	0.9998	0.9999							

TABLE A.2

t Distribution: Critical Values of *t*

		Significance level					
Degrees of freedom	*Two-tailed test*: *One-tailed test*:	10% 5%	5% 2.5%	2% 1%	1% 0.5%	0.2% 0.1%	0.1% 0.05%
1		6.314	12.706	31.821	63.657	318.309	636.619
2		2.920	4.303	6.965	9.925	22.327	31.599
3		2.353	3.182	4.541	5.841	10.215	12.924
4		2.132	2.776	3.747	4.604	7.173	8.610
5		2.015	2.571	3.365	4.032	5.893	6.869
6		1.943	2.447	3.143	3.707	5.208	5.959
7		1.894	2.365	2.998	3.499	4.785	5.408
8		1.860	2.306	2.896	3.355	4.501	5.041
9		1.833	2.262	2.821	3.250	4.297	4.781
10		1.812	2.228	2.764	3.169	4.144	4.587
11		1.796	2.201	2.718	3.106	4.025	4.437
12		1.782	2.179	2.681	3.055	3.930	4.318
13		1.771	2.160	2.650	3.012	3.852	4.221
14		1.761	2.145	2.624	2.977	3.787	4.140
15		1.753	2.131	2.602	2.947	3.733	4.073
16		1.746	2.120	2.583	2.921	3.686	4.015
17		1.740	2.110	2.567	2.898	3.646	3.965
18		1.734	2.101	2.552	2.878	3.610	3.922
19		1.729	2.093	2.539	2.861	3.579	3.883
20		1.725	2.086	2.528	2.845	3.552	3.850
21		1.721	2.080	2.518	2.831	3.527	3.819
22		1.717	2.074	2.508	2.819	3.505	3.792
23		1.714	2.069	2.500	2.807	3.485	3.768
24		1.711	2.064	2.492	2.797	3.467	3.745
25		1.708	2.060	2.485	2.787	3.450	3.725
26		1.706	2.056	2.479	2.779	3.435	3.707
27		1.703	2.052	2.473	2.771	3.421	3.690
28		1.701	2.048	2.467	2.763	3.408	3.674
29		1.699	2.045	2.462	2.756	3.396	3.659
30		1.697	2.042	2.457	2.750	3.385	3.646
32		1.694	2.037	2.449	2.738	3.365	3.622
34		1.691	2.032	2.441	2.728	3.348	3.601
36		1.688	2.028	2.434	2.719	3.333	3.582
38		1.686	2.024	2.429	2.712	3.319	3.566
40		1.684	2.021	2.423	2.704	3.307	3.551
42		1.682	2.018	2.418	2.698	3.296	3.538
44		1.680	2.015	2.414	2.692	3.286	3.526
46		1.679	2.013	2.410	2.687	3.277	3.515
48		1.677	2.011	2.407	2.682	3.269	3.505
50		1.676	2.009	2.403	2.678	3.261	3.496
60		1.671	2.000	2.390	2.660	3.232	3.460
70		1.667	1.994	2.381	2.648	3.211	3.435
80		1.664	1.990	2.374	2.639	3.195	3.416
90		1.662	1.987	2.368	2.632	3.183	3.402
100		1.660	1.984	2.364	2.626	3.174	3.390
120		1.658	1.980	2.358	2.617	3.160	3.373
150		1.655	1.976	2.351	2.609	3.145	3.357
200		1.653	1.972	2.345	2.601	3.131	3.340
300		1.650	1.968	2.339	2.592	3.118	3.323
400		1.649	1.966	2.336	2.588	3.111	3.315
500		1.648	1.965	2.334	2.586	3.107	3.310
600		1.647	1.964	2.333	2.584	3.104	3.307
∞		1.645	1.960	2.326	2.576	3.090	3.291

TABLE A.3

F Distribution: Critical Values of *F* (5% significance level)

ν_1 / ν_2	1	2	3	4	5	6	7	8	9	10	12	14	16	18	20
1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88	243.91	245.36	246.46	247.32	248.01
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.41	19.42	19.43	19.44	19.45
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.71	8.69	8.67	8.66
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.87	5.84	5.82	5.80
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.64	4.60	4.58	4.56
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.96	3.92	3.90	3.87
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.53	3.49	3.47	3.44
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.24	3.20	3.17	3.15
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.03	2.99	2.96	2.94
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.86	2.83	2.80	2.77
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.74	2.70	2.67	2.65
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.64	2.60	2.57	2.54
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.55	2.51	2.48	2.46
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.48	2.44	2.41	2.39
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.42	2.38	2.35	2.33
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.37	2.33	2.30	2.28
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.33	2.29	2.26	2.23
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.29	2.25	2.22	2.19
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.26	2.21	2.18	2.16
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.22	2.18	2.15	2.12
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.20	2.16	2.12	2.10
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.17	2.13	2.10	2.07
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.20	2.15	2.11	2.08	2.05
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.13	2.09	2.05	2.03
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	2.11	2.07	2.04	2.01
26	4.22	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.15	2.09	2.05	2.02	1.99
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20	2.13	2.08	2.04	2.00	1.97
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.12	2.06	2.02	1.99	1.96
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18	2.10	2.05	2.01	1.97	1.94
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.04	1.99	1.96	1.93
35	4.12	3.27	2.87	2.64	2.49	2.37	2.29	2.22	2.16	2.11	2.04	1.99	1.94	1.91	1.88
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.95	1.90	1.87	1.84
50	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07	2.03	1.95	1.89	1.85	1.81	1.78
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.86	1.82	1.78	1.75
70	3.98	3.13	2.74	2.50	2.35	2.23	2.14	2.07	2.02	1.97	1.89	1.84	1.79	1.75	1.72
80	3.96	3.11	2.72	2.49	2.33	2.21	2.13	2.06	2.00	1.95	1.88	1.82	1.77	1.73	1.70
90	3.95	3.10	2.71	2.47	2.32	2.20	2.11	2.04	1.99	1.94	1.86	1.80	1.76	1.72	1.69
100	3.94	3.09	2.70	2.46	2.31	2.19	2.10	2.03	1.97	1.93	1.85	1.79	1.75	1.71	1.68
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91	1.83	1.78	1.73	1.69	1.66
150	3.90	3.06	2.66	2.43	2.27	2.16	2.07	2.00	1.94	1.89	1.82	1.76	1.71	1.67	1.64
200	3.89	3.04	2.65	2.42	2.26	2.14	2.06	1.98	1.93	1.88	1.80	1.74	1.69	1.66	1.62
250	3.88	3.03	2.64	2.41	2.25	2.13	2.05	1.98	1.92	1.87	1.79	1.73	1.68	1.65	1.61
300	3.87	3.03	2.63	2.40	2.24	2.13	2.04	1.97	1.91	1.86	1.78	1.72	1.68	1.64	1.61
400	3.86	3.02	2.63	2.39	2.24	2.12	2.03	1.96	1.90	1.85	1.78	1.72	1.67	1.63	1.60
500	3.86	3.01	2.62	2.39	2.23	2.12	2.03	1.96	1.90	1.85	1.77	1.71	1.66	1.62	1.59
600	3.86	3.01	2.62	2.39	2.23	2.11	2.02	1.95	1.90	1.85	1.77	1.71	1.66	1.62	1.59
750	3.85	3.01	2.62	2.38	2.23	2.11	2.02	1.95	1.89	1.84	1.77	1.70	1.66	1.62	1.58
1000	3.85	3.00	2.61	2.38	2.22	2.11	2.02	1.95	1.89	1.84	1.76	1.70	1.65	1.61	1.58

TABLE A.3 (continued)

F Distribution: Critical Values of *F* (5% significance level)

v_1 / v_2	25	30	35	40	50	60	75	100	150	200
1	249.26	250.10	250.69	251.14	251.77	252.20	252.62	253.04	253.46	253.68
2	19.46	19.46	19.47	19.47	19.48	19.48	19.48	19.49	19.49	19.49
3	8.63	8.62	8.60	8.59	8.58	8.57	8.56	8.55	8.54	8.54
4	5.77	5.75	5.73	5.72	5.70	5.69	5.68	5.66	5.65	5.65
5	4.52	4.50	4.48	4.46	4.44	4.43	4.42	4.41	4.39	4.39
6	3.83	3.81	3.79	3.77	3.75	3.74	3.73	3.71	3.70	3.69
7	3.40	3.38	3.36	3.34	3.32	3.30	3.29	3.27	3.26	3.25
8	3.11	3.08	3.06	3.04	3.02	3.01	2.99	2.97	2.96	2.95
9	2.89	2.86	2.84	2.83	2.80	2.79	2.77	2.76	2.74	2.73
10	2.73	2.70	2.68	2.66	2.64	2.62	2.60	2.59	2.57	2.56
11	2.60	2.57	2.55	2.53	2.51	2.49	2.47	2.46	2.44	2.43
12	2.50	2.47	2.44	2.43	2.40	2.38	2.37	2.35	2.33	2.32
13	2.41	2.38	2.36	2.34	2.31	2.30	2.28	2.26	2.24	2.23
14	2.34	2.31	2.28	2.27	2.24	2.22	2.21	2.19	2.17	2.16
15	2.28	2.25	2.22	2.20	2.18	2.16	2.14	2.12	2.10	2.10
16	2.23	2.19	2.17	2.15	2.12	2.11	2.09	2.07	2.05	2.04
17	2.18	2.15	2.12	2.10	2.08	2.06	2.04	2.02	2.00	1.99
18	2.14	2.11	2.08	2.06	2.04	2.02	2.00	1.98	1.96	1.95
19	2.11	2.07	2.05	2.03	2.00	1.98	1.96	1.94	1.92	1.91
20	2.07	2.04	2.01	1.99	1.97	1.95	1.93	1.91	1.89	1.88
21	2.05	2.01	1.98	1.96	1.94	1.92	1.90	1.88	1.86	1.84
22	2.02	1.98	1.96	1.94	1.91	1.89	1.87	1.85	1.83	1.82
23	2.00	1.96	1.93	1.91	1.88	1.86	1.84	1.82	1.80	1.79
24	1.97	1.94	1.91	1.89	1.86	1.84	1.82	1.80	1.78	1.77
25	1.96	1.92	1.89	1.87	1.84	1.82	1.80	1.78	1.76	1.75
26	1.94	1.90	1.87	1.85	1.82	1.80	1.78	1.76	1.74	1.73
27	1.92	1.88	1.86	1.84	1.81	1.79	1.76	1.74	1.72	1.71
28	1.91	1.87	1.84	1.82	1.79	1.77	1.75	1.73	1.70	1.69
29	1.89	1.85	1.83	1.81	1.77	1.75	1.73	1.71	1.69	1.67
30	1.88	1.84	1.81	1.79	1.76	1.74	1.72	1.70	1.67	1.66
35	1.82	1.79	1.76	1.74	1.70	1.68	1.66	1.63	1.61	1.60
40	1.78	1.74	1.72	1.69	1.66	1.64	1.61	1.59	1.56	1.55
50	1.73	1.69	1.66	1.63	1.60	1.58	1.55	1.52	1.50	1.48
60	1.69	1.65	1.62	1.59	1.56	1.53	1.51	1.48	1.45	1.44
70	1.66	1.62	1.59	1.57	1.53	1.50	1.48	1.45	1.42	1.40
80	1.64	1.60	1.57	1.54	1.51	1.48	1.45	1.43	1.39	1.38
90	1.63	1.59	1.55	1.53	1.49	1.46	1.44	1.41	1.38	1.36
100	1.62	1.57	1.54	1.52	1.48	1.45	1.42	1.39	1.36	1.34
120	1.60	1.55	1.52	1.50	1.46	1.43	1.40	1.37	1.33	1.32
150	1.58	1.54	1.50	1.48	1.44	1.41	1.38	1.34	1.31	1.29
200	1.56	1.52	1.48	1.46	1.41	1.39	1.35	1.32	1.28	1.26
250	1.55	1.50	1.47	1.44	1.40	1.37	1.34	1.31	1.27	1.25
300	1.54	1.50	1.46	1.43	1.39	1.36	1.33	1.30	1.26	1.23
400	1.53	1.49	1.45	1.42	1.38	1.35	1.32	1.28	1.24	1.22
500	1.53	1.48	1.45	1.42	1.38	1.35	1.31	1.28	1.23	1.21
600	1.52	1.48	1.44	1.41	1.37	1.34	1.31	1.27	1.23	1.20
750	1.52	1.47	1.44	1.41	1.37	1.34	1.30	1.26	1.22	1.20
1000	1.52	1.47	1.43	1.41	1.36	1.33	1.30	1.26	1.22	1.19

.

TABLE A.3 (continued)

F Distribution: Critical Values of *F* (1% significance level)

v_2 \ v_1	1	2	3	4	5	6	7	8	9	10	12	14	16	18	20
1	4052.18	4999.50	5403.35	5624.58	5763.65	5858.99	5928.36	5981.07	6022.47	6055.85	6106.32	6142.67	6170.10	6191.53	6208.73
2	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39	99.40	99.42	99.43	99.44	99.44	99.45
3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35	27.23	27.05	26.92	26.83	26.75	26.69
4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55	14.37	14.25	14.15	14.08	14.02
5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05	9.89	9.77	9.68	9.61	9.55
6	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.72	7.60	7.52	7.45	7.40
7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.47	6.36	6.28	6.21	6.16
8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.67	5.56	5.48	5.41	5.36
9	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.11	5.01	4.92	4.86	4.81
10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.71	4.60	4.52	4.46	4.41
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.40	4.29	4.21	4.15	4.10
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.16	4.05	3.97	3.91	3.86
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	3.96	3.86	3.78	3.72	3.66
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94	3.80	3.70	3.62	3.56	3.51
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.67	3.56	3.49	3.42	3.37
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.55	3.45	3.37	3.31	3.26
17	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.46	3.35	3.27	3.21	3.16
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.37	3.27	3.19	3.13	3.08
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.30	3.19	3.12	3.05	3.00
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.23	3.13	3.05	2.99	2.94
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31	3.17	3.07	2.99	2.93	2.88
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.12	3.02	2.94	2.88	2.83
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21	3.07	2.97	2.89	2.83	2.78
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	3.03	2.93	2.85	2.79	2.74
25	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22	3.13	2.99	2.89	2.81	2.75	2.70
26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09	2.96	2.86	2.78	2.72	2.66
27	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15	3.06	2.93	2.82	2.75	2.68	2.63
28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03	2.90	2.79	2.72	2.65	2.60
29	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09	3.00	2.87	2.77	2.69	2.63	2.57
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.84	2.74	2.66	2.60	2.55
35	7.42	5.27	4.40	3.91	3.59	3.37	3.20	3.07	2.96	2.88	2.74	2.64	2.56	2.50	2.44
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.66	2.56	2.48	2.42	2.37
50	7.17	5.06	4.20	3.72	3.41	3.19	3.02	2.89	2.78	2.70	2.56	2.46	2.38	2.32	2.27
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.50	2.39	2.31	2.25	2.20
70	7.01	4.92	4.07	3.60	3.29	3.07	2.91	2.78	2.67	2.59	2.45	2.35	2.27	2.20	2.15
80	6.96	4.88	4.04	3.56	3.26	3.04	2.87	2.74	2.64	2.55	2.42	2.31	2.23	2.17	2.12
90	6.93	4.85	4.01	3.53	3.23	3.01	2.84	2.72	2.61	2.52	2.39	2.29	2.21	2.14	2.09
100	6.90	4.82	3.98	3.51	3.21	2.99	2.82	2.69	2.59	2.50	2.37	2.27	2.19	2.12	2.07
120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47	2.34	2.23	2.15	2.09	2.03
150	6.81	4.75	3.91	3.45	3.14	2.92	2.76	2.63	2.53	2.44	2.31	2.20	2.12	2.06	2.00
200	6.76	4.71	3.88	3.41	3.11	2.89	2.73	2.60	2.50	2.41	2.27	2.17	2.09	2.03	1.97
250	6.74	4.69	3.86	3.40	3.09	2.87	2.71	2.58	2.48	2.39	2.26	2.15	2.07	2.01	1.95
300	6.72	4.68	3.85	3.38	3.08	2.86	2.70	2.57	2.47	2.38	2.24	2.14	2.06	1.99	1.94
400	6.70	4.66	3.83	3.37	3.06	2.85	2.68	2.56	2.45	2.37	2.23	2.13	2.05	1.98	1.92
500	6.69	4.65	3.82	3.36	3.05	2.84	2.68	2.55	2.44	2.36	2.22	2.12	2.04	1.97	1.92
600	6.68	4.64	3.81	3.35	3.05	2.83	2.67	2.54	2.44	2.35	2.21	2.11	2.03	1.96	1.91
750	6.67	4.63	3.81	3.34	3.04	2.83	2.66	2.53	2.43	2.34	2.21	2.11	2.02	1.96	1.90
1000	6.66	4.63	3.80	3.34	3.04	2.82	2.66	2.53	2.43	2.34	2.20	2.10	2.02	1.95	1.90

TABLE A.3 (continued)

F Distribution: Critical Values of *F* (1% significance level)

v_1 / v_2	25	30	35	40	50	60	75	100	150	200
1	6239.83	6260.65	6275.57	6286.78	6302.52	6313.03	6323.56	6334.11	6344.68	6349.97
2	99.46	99.47	99.47	99.47	99.48	99.48	99.49	99.49	99.49	99.49
3	26.58	26.50	26.45	26.41	26.35	26.32	26.28	26.24	26.20	26.18
4	13.91	13.84	13.79	13.75	13.69	13.65	13.61	13.58	13.54	13.52
5	9.45	9.38	9.33	9.29	9.24	9.20	9.17	9.13	9.09	9.08
6	7.30	7.23	7.18	7.14	7.09	7.06	7.02	6.99	6.95	6.93
7	6.06	5.99	5.94	5.91	5.86	5.82	5.79	5.75	5.72	5.70
8	5.26	5.20	5.15	5.12	5.07	5.03	5.00	4.96	4.93	4.91
9	4.71	4.65	4.60	4.57	4.52	4.48	4.45	4.41	4.38	4.36
10	4.31	4.25	4.20	4.17	4.12	4.08	4.05	4.01	3.98	3.96
11	4.01	3.94	3.89	3.86	3.81	3.78	3.74	3.71	3.67	3.66
12	3.76	3.70	3.65	3.62	3.57	3.54	3.50	3.47	3.43	3.41
13	3.57	3.51	3.46	3.43	3.38	3.34	3.31	3.27	3.24	3.22
14	3.41	3.35	3.30	3.27	3.22	3.18	3.15	3.11	3.08	3.06
15	3.28	3.21	3.17	3.13	3.08	3.05	3.01	2.98	2.94	2.92
16	3.16	3.10	3.05	3.02	2.97	2.93	2.90	2.86	2.83	2.81
17	3.07	3.00	2.96	2.92	2.87	2.83	2.80	2.76	2.73	2.71
18	2.98	2.92	2.87	2.84	2.78	2.75	2.71	2.68	2.64	2.62
19	2.91	2.84	2.80	2.76	2.71	2.67	2.64	2.60	2.57	2.55
20	2.84	2.78	2.73	2.69	2.64	2.61	2.57	2.54	2.50	2.48
21	2.79	2.72	2.67	2.64	2.58	2.55	2.51	2.48	2.44	2.42
22	2.73	2.67	2.62	2.58	2.53	2.50	2.46	2.42	2.38	2.36
23	2.69	2.62	2.57	2.54	2.48	2.45	2.41	2.37	2.34	2.32
24	2.64	2.58	2.53	2.49	2.44	2.40	2.37	2.33	2.29	2.27
25	2.60	2.54	2.49	2.45	2.40	2.36	2.33	2.29	2.25	2.23
26	2.57	2.50	2.45	2.42	2.36	2.33	2.29	2.25	2.21	2.19
27	2.54	2.47	2.42	2.38	2.33	2.29	2.26	2.22	2.18	2.16
28	2.51	2.44	2.39	2.35	2.30	2.26	2.23	2.19	2.15	2.13
29	2.48	2.41	2.36	2.33	2.27	2.23	2.20	2.16	2.12	2.10
30	2.45	2.39	2.34	2.30	2.25	2.21	2.17	2.13	2.09	2.07
35	2.35	2.28	2.23	2.19	2.14	2.10	2.06	2.02	1.98	1.96
40	2.27	2.20	2.15	2.11	2.06	2.02	1.98	1.94	1.90	1.87
50	2.17	2.10	2.05	2.01	1.95	1.91	1.87	1.82	1.78	1.76
60	2.10	2.03	1.98	1.94	1.88	1.84	1.79	1.75	1.70	1.68
70	2.05	1.98	1.93	1.89	1.83	1.78	1.74	1.70	1.65	1.62
80	2.01	1.94	1.89	1.85	1.79	1.75	1.70	1.65	1.61	1.58
90	1.99	1.92	1.86	1.82	1.76	1.72	1.67	1.62	1.57	1.55
100	1.97	1.89	1.84	1.80	1.74	1.69	1.65	1.60	1.55	1.52
120	1.93	1.86	1.81	1.76	1.70	1.66	1.61	1.56	1.51	1.48
150	1.90	1.83	1.77	1.73	1.66	1.62	1.57	1.52	1.46	1.43
200	1.87	1.79	1.74	1.69	1.63	1.58	1.53	1.48	1.42	1.39
250	1.85	1.77	1.72	1.67	1.61	1.56	1.51	1.46	1.40	1.36
300	1.84	1.76	1.70	1.66	1.59	1.55	1.50	1.44	1.38	1.35
400	1.82	1.75	1.69	1.64	1.58	1.53	1.48	1.42	1.36	1.32
500	1.81	1.74	1.68	1.63	1.57	1.52	1.47	1.41	1.34	1.31
600	1.80	1.73	1.67	1.63	1.56	1.51	1.46	1.40	1.34	1.30
750	1.80	1.72	1.66	1.62	1.55	1.50	1.45	1.39	1.33	1.29
1000	1.79	1.72	1.66	1.61	1.54	1.50	1.44	1.38	1.32	1.28

.

TABLE A.3 (continued)

F Distribution: Critical Values of *F* (0.1% significance level)

v_2 \ v_1	1	2	3	4	5	6	7	8	9	10	12	14	16	18	20
1	4.05e05	5.00e05	5.40e05	5.62e05	5.76e05	5.86e05	5.93e05	5.98e05	6.02e05	6.06e05	6.11e05	6.14e05	6.17e05	6.19e05	6.21e05
2	998.50	999.00	999.17	999.25	999.30	999.33	999.36	999.37	999.39	999.40	999.42	999.43	999.44	999.44	999.45
3	167.03	148.50	141.11	137.10	134.58	132.85	131.58	130.62	129.86	129.25	128.32	127.64	127.14	126.74	126.42
4	74.14	61.25	56.18	53.44	51.71	50.53	49.66	49.00	48.47	48.05	47.41	46.95	46.60	46.32	46.10
5	47.18	37.12	33.20	31.09	29.75	28.83	28.16	27.65	27.24	26.92	26.42	26.06	25.78	25.57	25.39
6	35.51	27.00	23.70	21.92	20.80	20.03	19.46	19.03	18.69	18.41	17.99	17.68	17.45	17.27	17.12
7	29.25	21.69	18.77	17.20	16.21	15.52	15.02	14.63	14.33	14.08	13.71	13.43	13.23	13.06	12.93
8	25.41	18.49	15.83	14.39	13.48	12.86	12.40	12.05	11.77	11.54	11.19	10.94	10.75	10.60	10.48
9	22.86	16.39	13.90	12.56	11.71	11.13	10.70	10.37	10.11	9.89	9.57	9.33	9.15	9.01	8.90
10	21.04	14.91	12.55	11.28	10.48	9.93	9.52	9.20	8.96	8.75	8.45	8.22	8.05	7.91	7.80
11	19.69	13.81	11.56	10.35	9.58	9.05	8.66	8.35	8.12	7.92	7.63	7.41	7.24	7.11	7.01
12	18.64	12.97	10.80	9.63	8.89	8.38	8.00	7.71	7.48	7.29	7.00	6.79	6.63	6.51	6.40
13	17.82	12.31	10.21	9.07	8.35	7.86	7.49	7.21	6.98	6.80	6.52	6.31	6.16	6.03	5.93
14	17.14	11.78	9.73	8.62	7.92	7.44	7.08	6.80	6.58	6.40	6.13	5.93	5.78	5.66	5.56
15	16.59	11.34	9.34	8.25	7.57	7.09	6.74	6.47	6.26	6.08	5.81	5.62	5.46	5.35	5.25
16	16.12	10.97	9.01	7.94	7.27	6.80	6.46	6.19	5.98	5.81	5.55	5.35	5.20	5.09	4.99
17	15.72	10.66	8.73	7.68	7.02	6.56	6.22	5.96	5.75	5.58	5.32	5.13	4.99	4.87	4.78
18	15.38	10.39	8.49	7.46	6.81	6.35	6.02	5.76	5.56	5.39	5.13	4.94	4.80	4.68	4.59
19	15.08	10.16	8.28	7.27	6.62	6.18	5.85	5.59	5.39	5.22	4.97	4.78	4.64	4.52	4.43
20	14.82	9.95	8.10	7.10	6.46	6.02	5.69	5.44	5.24	5.08	4.82	4.64	4.49	4.38	4.29
21	14.59	9.77	7.94	6.95	6.32	5.88	5.56	5.31	5.11	4.95	4.70	4.51	4.37	4.26	4.17
22	14.38	9.61	7.80	6.81	6.19	5.76	5.44	5.19	4.99	4.83	4.58	4.40	4.26	4.15	4.06
23	14.20	9.47	7.67	6.70	6.08	5.65	5.33	5.09	4.89	4.73	4.48	4.30	4.16	4.05	3.96
24	14.03	9.34	7.55	6.59	5.98	5.55	5.23	4.99	4.80	4.64	4.39	4.21	4.07	3.96	3.87
25	13.88	9.22	7.45	6.49	5.89	5.46	5.15	4.91	4.71	4.56	4.31	4.13	3.99	3.88	3.79
26	13.74	9.12	7.36	6.41	5.80	5.38	5.07	4.83	4.64	4.48	4.24	4.06	3.92	3.81	3.72
27	13.61	9.02	7.27	6.33	5.73	5.31	5.00	4.76	4.57	4.41	4.17	3.99	3.86	3.75	3.66
28	13.50	8.93	7.19	6.25	5.66	5.24	4.93	4.69	4.50	4.35	4.11	3.93	3.80	3.69	3.60
29	13.39	8.85	7.12	6.19	5.59	5.18	4.87	4.64	4.45	4.29	4.05	3.88	3.74	3.63	3.54
30	13.29	8.77	7.05	6.12	5.53	5.12	4.82	4.58	4.39	4.24	4.00	3.82	3.69	3.58	3.49
35	12.90	8.47	6.79	5.88	5.30	4.89	4.59	4.36	4.18	4.03	3.79	3.62	3.48	3.38	3.29
40	12.61	8.25	6.59	5.70	5.13	4.73	4.44	4.21	4.02	3.87	3.64	3.47	3.34	3.23	3.14
50	12.22	7.96	6.34	5.46	4.90	4.51	4.22	4.00	3.82	3.67	3.44	3.27	3.41	3.04	2.95
60	11.97	7.77	6.17	5.31	4.76	4.37	4.09	3.86	3.69	3.54	3.32	3.15	3.02	2.91	2.83
70	11.80	7.64	6.06	5.20	4.66	4.28	3.99	3.77	3.60	3.45	3.23	3.06	2.93	2.83	2.74
80	11.67	7.54	5.97	5.12	4.58	4.20	3.92	3.70	3.53	3.39	3.16	3.00	2.87	2.76	2.68
90	11.57	7.47	5.91	5.06	4.53	4.15	3.87	3.65	3.48	3.34	3.11	2.95	2.82	2.71	2.63
100	11.50	7.41	5.86	5.02	4.48	4.11	3.83	3.61	3.44	3.30	3.07	2.91	2.78	2.68	2.59
120	11.38	7.32	5.78	4.95	4.42	4.04	3.77	3.55	3.38	3.24	3.02	2.85	2.72	2.62	2.53
150	11.27	7.24	5.71	4.88	4.35	3.98	3.71	3.49	3.32	3.18	2.96	2.80	2.67	2.56	2.48
200	11.15	7.15	5.63	4.81	4.29	3.92	3.65	3.43	3.26	3.12	2.90	2.74	2.61	2.51	2.42
250	11.09	7.10	5.59	4.77	4.25	3.88	3.61	3.40	3.23	3.09	2.87	2.71	2.58	2.48	2.39
300	11.04	7.07	5.56	4.75	4.22	3.86	3.59	3.38	3.21	3.07	2.85	2.69	2.56	2.46	2.37
400	10.99	7.03	5.53	4.71	4.19	3.83	3.56	3.35	3.18	3.04	2.82	2.66	2.53	2.43	2.34
500	10.96	7.00	5.51	4.69	4.18	3.81	3.54	3.33	3.16	3.02	2.81	2.64	2.52	2.41	2.33
600	10.94	6.99	5.49	4.68	4.16	3.80	3.53	3.32	3.15	3.01	2.80	2.63	2.51	2.40	2.32
750	10.91	6.97	5.48	4.67	4.15	3.79	3.52	3.31	3.14	3.00	2.78	2.62	2.49	2.39	2.31
1000	10.89	6.96	5.46	4.65	4.14	3.78	3.51	3.30	3.13	2.99	2.77	2.61	2.48	2.38	2.30

.

TABLE A.3 (continued)

F Distribution: Critical Values of *F* (0.1% significance level)

v_1 / v_2	25	30	35	40	50	60	75	100	150	200
1	6.24e05	6.26e05	6.28e05	6.29e05	6.30e05	6.31e05	6.32e05	6.33e05	6.35e05	6.35e05
2	999.46	999.47	999.47	999.47	999.48	999.48	999.49	999.49	999.49	999.49
3	125.84	125.45	125.17	124.96	124.66	124.47	124.27	124.07	123.87	123.77
4	45.70	45.43	45.23	45.09	44.88	44.75	44.61	44.47	44.33	44.26
5	25.08	24.87	24.72	24.60	24.44	24.33	24.22	24.12	24.01	23.95
6	16.85	16.67	16.54	16.44	16.31	16.21	16.12	16.03	15.93	15.89
7	12.69	12.53	12.41	12.33	12.20	12.12	12.04	11.95	11.87	11.82
8	10.26	10.11	10.00	9.92	9.80	9.73	9.65	9.57	9.49	9.45
9	8.69	8.55	8.46	8.37	8.26	8.19	8.11	8.04	7.96	7.93
10	7.60	7.47	7.37	7.30	7.19	7.12	7.05	6.98	6.91	6.87
11	6.81	6.68	6.59	6.52	6.42	6.35	6.28	6.21	6.14	6.10
12	6.22	6.09	6.00	5.93	5.83	5.76	5.70	5.63	5.56	5.52
13	5.75	5.63	5.54	5.47	5.37	5.30	5.24	5.17	5.10	5.07
14	5.38	5.25	5.17	5.10	5.00	4.94	4.87	4.81	4.74	4.71
15	5.07	4.95	4.86	4.80	4.70	4.64	4.57	4.51	4.44	4.41
16	4.82	4.70	4.61	4.54	4.45	4.39	4.32	4.26	4.19	4.16
17	4.60	4.48	4.40	4.33	4.24	4.18	4.11	4.05	3.98	3.95
18	4.42	4.30	4.22	4.15	4.06	4.00	3.93	3.87	3.80	3.77
19	4.26	4.14	4.06	3.99	3.90	3.84	3.78	3.71	3.65	3.61
20	4.12	4.00	3.92	3.86	3.77	3.70	3.64	3.58	3.51	3.48
21	4.00	3.88	3.80	3.74	3.64	3.58	3.52	3.46	3.39	3.36
22	3.89	3.78	3.70	3.63	3.54	3.48	3.41	3.35	3.28	3.25
23	3.79	3.68	3.60	3.53	3.44	3.38	3.32	3.25	3.19	3.16
24	3.71	3.59	3.51	3.45	3.36	3.29	3.23	3.17	3.10	3.07
25	3.63	3.52	3.43	3.37	3.28	3.22	3.15	3.09	3.03	2.99
26	3.56	3.44	3.36	3.30	3.21	3.15	3.08	3.02	2.95	2.92
27	3.49	3.38	3.30	3.23	3.14	3.08	3.02	2.96	2.89	2.86
28	3.43	3.32	3.24	3.18	3.09	3.02	2.96	2.90	2.83	2.80
29	3.38	3.27	3.18	3.12	3.03	2.97	2.91	2.84	2.78	2.74
30	3.33	3.22	3.13	3.07	2.98	2.92	2.86	2.79	2.73	2.69
35	3.13	3.02	2.93	2.87	2.78	2.72	2.66	2.59	2.52	2.49
40	2.98	2.87	2.79	2.73	2.64	2.57	2.51	2.44	2.38	2.34
50	2.79	2.68	2.60	2.53	2.44	2.38	2.31	2.25	2.18	2.14
60	2.67	2.55	2.47	2.41	2.32	2.25	2.19	2.12	2.05	2.01
70	2.58	2.47	2.39	2.32	2.23	2.16	2.10	2.03	1.95	1.92
80	2.52	2.41	2.32	2.26	2.16	2.10	2.03	1.96	1.89	1.85
90	2.47	2.36	2.27	2.21	2.11	2.05	1.98	1.91	1.83	1.79
100	2.43	2.32	2.24	2.17	2.08	2.01	1.94	1.87	1.79	1.75
120	2.37	2.26	2.18	2.11	2.02	1.95	1.88	1.81	1.73	1.68
150	2.32	2.21	2.12	2.06	1.96	1.89	1.82	1.74	1.66	1.62
200	2.26	2.15	2.07	2.00	1.90	1.83	1.76	1.68	1.60	1.55
250	2.23	2.12	2.03	1.97	1.87	1.80	1.72	1.65	1.56	1.51
300	2.21	2.10	2.01	1.94	1.85	1.78	1.70	1.62	1.53	1.48
400	2.18	2.07	1.98	1.92	1.82	1.75	1.67	1.59	1.50	1.45
500	2.17	2.05	1.97	1.90	1.80	1.73	1.65	1.57	1.48	1.43
600	2.16	2.04	1.96	1.89	1.79	1.72	1.64	1.56	1.46	1.41
750	2.15	2.03	1.95	1.88	1.78	1.71	1.63	1.55	1.45	1.40
1000	2.14	2.02	1.94	1.87	1.77	1.69	1.62	1.53	1.44	1.38

.

TABLE A.4

χ^2 (Chi-Squared) Distribution: Critical Values of χ^2

	Significance level		
Degrees of freedom	5%	1%	0.1%
1	3.841	6.635	10.828
2	5.991	9.210	13.816
3	7.815	11.345	16.266
4	9.488	13.277	18.467
5	11.070	15.086	20.515
6	12.592	16.812	22.458
7	14.067	18.475	24.322
8	15.507	20.090	26.124
9	16.919	21.666	27.877
10	18.307	23.209	29.588

.