PROBABILITY THEORY, STATISTICAL INFERENCE, AND STOCHASTIC PROCESSES

DR N L BHIKSHU | DR G MAHESH
DR MD YOUNUS | MR T VAMSIDAR GOUD
MR K RAGHURAM BHATTAR

BlueRose ONE.com
Stories Matter
NewDelhi • London

BLUEROSE PUBLISHERS
India | U.K.

BlueRoseONE.com
Stories Matter
NewDelhi • London

For permissions requests or inquiries regarding this publication,
please contact:

BLUEROSE PUBLISHERS
www.BlueRoseONE.com
info@bluerosepublishers.com
+91 8882 898 898
+4407342408967

ISBN: 978-93-7825-952-4

Cover Design: Ajay Dhyani

First Edition: September 2026

Syllabus

Unit-1: Probability Basics and Baye's theorem: Random Experiment, sample space, events, probability, axioms of probability, some elementary theorems, Addition theorem, Multiplication theorem, Conditional probability, independent and dependent events, Baye's theorem.

Unit-2: Random Variables: Types of Random Variables, Probability Distribution and probability function of discrete random variable, Distribution function of discrete random variables, Measure of a discrete random variable, Mean and variance of linear combination of Random variables, Probability Distribution and probability function of continuous random variable, Distribution function of continuous random variables, Measure of a continuous random variable, Joint probability functions of discrete and continuous random variables.

Unit-3: Discrete Distributions: Binomial distributions, measures of binomial distributions, recurrence relation, fitting a binomial distribution, Chebshev's theorem, Poisson distributions, measures of Poisson distributions, Fitting a Poisson distribution. Poisson process.

Continuous Distributions: Normal distribution, measures of normal distribution, normal approximation to binomial distribution. Uniform distribution, measures of Uniform distribution.

Unit-4: Sampling distribution: sampling methods (probability and non-probability), classification of samples, sampling distribution of mean, central limit theorem. Standard error, finite population. Estimation, point and interval estimation, confidence intervals. Maximum Likelihood Estimation (MLE).

Unit-5: Testing of Hypothesis (Large Sample): Introduction, null hypothesis, alternate hypothesis, errors in decision making, testing of Hypothesis for single mean, difference of two means, single proportion, difference of two proportions.

Unit-6: Testing of Hypothesis (Small Sample): Student's 't' test, testing of Hypothesis for single mean, difference of two means, F- test for difference between two variances, test for equality between several means, χ^2- test for population variance, χ^2- test for goodness of fit, χ^2- test for independence of attributes.

Unit-7: Introduction to Correlation, Types of Correlation, Methods of Studying Correlation, Karl Pearson’s Correlation Coefficient, Spearman’s Rank Correlation.

Introduction to Regression, Regression Lines, Method of Least Squares: Straight Line, Parabola, Exponential Curve, Power Curve, Logarithmic Curve, multiple linear regression, logistic regression.

Unit-8: Stochastic Process: Introduction to Stochastic Processes- Markov process. Transition Probability, Transition Probability Matrix, First order and Higher order Markov process, n – step transition probabilities, Markov Chain, Steady state condition, Markov analysis.

Contents

UNIT 1

Probability

Introduction

We often come across statements like "There is a chance that the team will win the match," or "The stock market may rise tomorrow." “It might rain today,” "The train is likely to arrive late,". In these situations, we can estimate the likelihood of each possibility, but we are unable to predict the exact outcome with certainty. A mathematical method for measuring this kind of uncertainty is provided by probability, which helps in decision-making. Before we formally define probability, it is helpful to understand some basic concepts related to it.

Addition Rule (Sum Rule):

If one event occurs in m different ways and another event occurs in n different ways and if both cannot occur simultaneously, then the total number of ways of occurring of one of the events is m + n ways.

Ex: If there are 3 topics in Maths, 4 topics in BEE, 5 topics in Physics for a student to do a project, then the number of ways of selecting a topic for doing a project is 3+4+5 = 12 ways.

Multiplication Rule (Product Rule): If a task can be done in m ways, and for each of these, a second task can be done in n ways, then the total number of ways of performing both the tasks together is **m x n** ways. This principle can be extended to more than two tasks.

Ex: Suppose a person wants to go to a party. He has 3 pants, 4 shirts and 2 ties. The number of ways of selecting a pant, a shirt and a tie to wear for a party is 3 x 4 x 2 = 24

Permutation: From a given finite set of objects or elements, selecting some or all of them and arranging them linearly is called linear permutation or permutation and arranging them round a circle is called circular permutation.

Number of permutations of n different objects taken r at a time is given by

$$\boldsymbol{n_{p_r}} = \frac{\boldsymbol{n}!}{(\boldsymbol{n}-\boldsymbol{r})!}$$

Number of permutations of n different objects taken all a time is $\boldsymbol{n_{p_n}} = n!$

Number of permutations of n different objects taken r at a time when repetitions are allowed is n^r

Number of permutations of n different objects taken r at a time with at least one repetition is $n^r - \boldsymbol{n_{p_r}}$

Number of permutations of n objects of which m objects are of one kind, k objects are of second kind is $\frac{n!}{k!\,.m!}$

Number of circular permutations of n objects taken all at a time is $(n-1)!$ and r at a time is $\frac{n_{p_r}}{r}$

Number of hanging type circular permutations of n objects taken all at a time is $\frac{(n-1)!}{2}$ and r at a time is $\frac{n_{p_r}}{2r}$.

Combination: Number of ways of selecting r objects from n objects without considering the order is called combination. It is given by $\boldsymbol{n_{c_r}} = \frac{n!}{(n-r)!\,r!}$

Note: Permutation is the different ways of arranging a set of objects where the sequence or order is important. Whereas combination is a selection of objects where the order does NOT matter.

A deterministic experiment is an experiment whose result is unique and can be predicted with certainty prior to the conduction of an experiment.

A probabilistic or non-deterministic or random experiment: An experiment whose result is not unique and cannot be predicted with certainty even though the outcomes of this experiment are known in advance is called a Random Experiment.

Ex: If the coin is tossed, outcomes are head and tail, but cannot be predicted whether it would be tail or head at that moment. Such an experiment is called a random experiment.

Trial: A single performance of an experiment is called a trial.

Sample space: The set of all possible outcomes of a random experiment is called the sample space. It is denoted by S

Ex: When a die is tossed sample space is $S = \{1,2,3,4,5,6\}$

Event: A finite subset of a sample space is called an Event.

An event containing only one outcome is called an **elementary event** and containing more than one is called a **compound event.**

Exhaustive events are a set of all possible outcomes of a random experiment. Exhaustive events are a collection of events such that together they include all possible outcomes of the experiment.

Throwing a die:

- Event A = Even numbers = $\{2,4,6\}$
- Event B = Odd numbers = $\{1,3,5\}$

A and B together form exhaustive events because they cover all outcomes.

Mutually Exclusive Events: Two or more events are said to be mutually exclusive events if the occurrence of one of the events prevents the occurrence of any of the remaining events.

Ex: In tossing a coin, the events heads and tails are mutually exclusive

Equally likely events: Two or more events are said to be equally likely if there is no reason to expect one in preference to others.

Equally likely events are events that have the same chance (probability) of occurring.

Ex: In tossing a coin, getting head or tail are equally likely

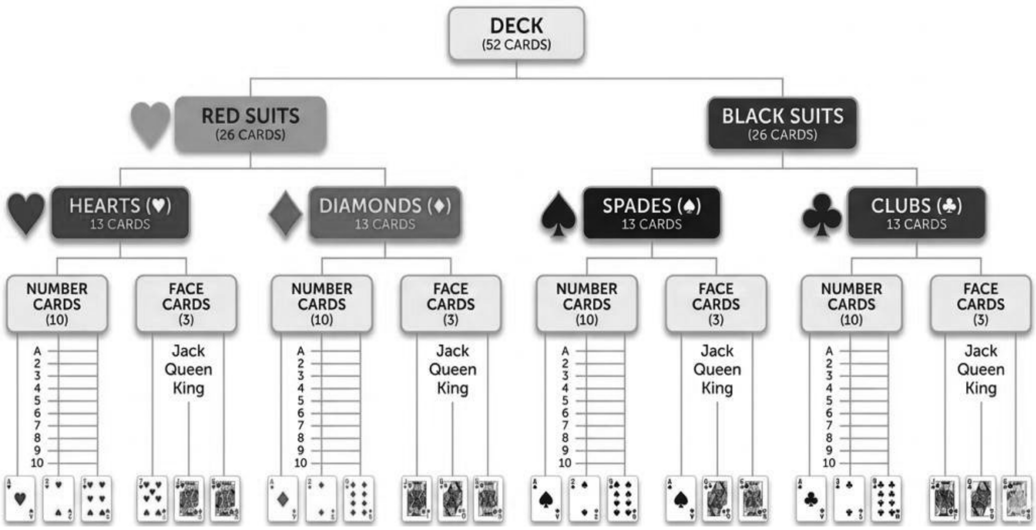
Certain or Sure Event: An event which always happens is called a Sure event.

Ex: In tossing a coin, getting head or tail is a sure event

Impossible or Null Event: An event that never happens is called an impossible event.

Ex: In tossing a coin, getting head and tail is an impossible event.

The structure of standard deck of playing cards



1. Classical Definition of Probability: If the random experiment results in n exhaustive, mutually exclusive and equally likely events, and m of them are favourable for happening of an event A, then the probability of the happening of event A is

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

$n - m$ are not favourable for happening of event A so Probability of non happening of an event A is given by $P(A^C) = \frac{n-m}{n} = 1 - \frac{m}{n} = 1 - P(A)$

Note: i. This definition will fail if number of outcomes are infinite and not equally likely

ii. In probability, for “and” we use multiplication and for “or” we use addition.

2. Empirical Definition of Probability: If the event A happens m times out of n trials, then the probability (p) of happening of an event A is given by p = $P(A) = \lim_{n \to \infty} \frac{m}{n}$

Note: P(Sure event) = 1

Note: P(impossible event) = 0

Axioms of Probability:

There are 3 axioms of probability

1. If A is an event of a sample space S, then the probability of that event lies between 0 and 1 i.e. 0$\leq P(A) \leq 1$
2. $P(S) = 1$
3. If A and B are any two mutually exclusive events of the sample space, then

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

   Simple Probability

	March	Not-March	Total
Friday	4	48	52
Non-Friday	27	286	313
Total	31	334	365

P(March) = $\frac{31}{365}$, P(Friday) = $\frac{52}{365}$

Solved Problems

1. Find the probability of getting no head in tossing two coins

Sol. when two coins are tossed, sample space S = {HH, HT, TH, TT}

Total number of cases, n = 4 and of them one is favourable, so m = 1

So P(no heads) = $\frac{m}{n} = \frac{1}{4}$

2. Find the probability of getting at least one head in tossing 3 coins

Sol. When three coins are tossed, sample space S =

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

Total number of cases, n = 8, and 7 of them are favourable, so m = 7

P(At least one head) = $\frac{m}{n} = \frac{7}{8}$

(OR) P(At least one head) = 1-P(getting no head) = $1 - \frac{1}{8} = \frac{7}{8}$

3.What is the probability that a leap year will contain 53 Mondays and 53 Tuesdays?

Sol. A leap year has 366 days = 52 weeks + 2 days.

Those 2 days can be in any way like {SM, MT, TW, WT, TF, FS, SS}

So total number of cases are 7 and only one is favourable for getting 53 Mondays and 53 Tuesdays i.e. {MT}

So required probability = 1/7

4. What is the probability of getting i) a red card ii) Ace iii) Kings from a pack of cards?

Sol. Number of cards in a pack of cards n = 52

i) Number of red card m = 26

So P(Red cards) = $\frac{m}{n} = \frac{26}{52} = \frac{1}{2}$

ii) Number of Aces = m = 4

So P(Ace card) = $\frac{4}{52} = \frac{1}{13}$

iii) Number of kings = m = 4

So P(King) = $\frac{4}{52} = \frac{1}{13}$

5. A die is thrown. Find the probability of getting:

a) Even number

b) Number greater than 4

Sol. Sample space S = {1, 2, 3, 4, 5, 6}

Here, n = 6

a) m = 3, since {2, 4, 6}

So, P(getting even number) = 3/6 = 1/2

b) m = 2, since {5, 6}

So, P(getting Number greater than 4) = 2/6 = 1/3

6. A batch has 10 items, 2 are defective. One item is chosen randomly. Find the probability that it is defective?

Sol. Total no of items are 10.

Let D represents defective items, then P(D) = $\frac{2}{10} = \frac{1}{5}$

7. An electronic device works with probability 0.8. Find the probability it fails.

Sol. P(failure) = 1− 0.8 = 0.2

8. To m letters there are m addressed envelopes. These letters are put in envelopes at random.

i) What is the probability that all the letters are in right addressed envelopes?

ii) What is the probability that at least one letter will go into an incorrectly addressed envelope?

Sol. Total no of letters and addressed envelopes are m. Some letters can be placed in

m addressed envelope in m! ways.

i) But in only one way all the m letters will go into m right addressed envelopes

So, Probability = $\frac{1}{m!}$

ii) P(At least one letter going into incorrect addressed envelope) = 1- P(none going into incorrect addressed envelope) = $1 - \frac{1}{m!}$

9. There are 15 objects in a bag, out of which 6 are not in a good condition. If 5 objects are selected at random, what is the probability that i) no object is good ii) only 3 are not good?

Sol. Total number of items in a bag = 15, 6 defective and 9 non-defective (good).

i) P(no object is good) = $\frac{6_{c_5}}{15_{c_5}}$ (Since selecting 5 out of 6 defective items is 6_{c_5})

ii) P(3 are not good) = P(3 are not good and remaining 2 are good) = $\frac{6_{c_3} 9_{c_2}}{15_{c_5}}$ (since selecting 3 out of 6 defectives and 2 out of 9 good)

10. In a class there are 12 boys and 6 girls. A group of 5 students is to be formed from that class. Find the probability for the group to contain at least 3 girls?

Sol. A group of 5 students can be formed from 12+ 6 = 18 students in 18_{c_5} ways

i.e $n(S) = 18_{c_5} = 8598$

Let A be the event of forming a group with at least 3 girls. So the group can have 3 girls and 2 boys or 4 girls and 1 boy or 5 girls

So total no of ways of forming a committee $n(A) = 6_{c_3}$ x 12_{c_2}+ 6_{c_4}x 12_{c_1}+ 6_{c_5} = 1506

$$P(A) = \frac{n(A)}{n(S)} = \frac{1506}{8568} = 0.175$$

11. Find the probability of getting sum as 10 if we throw two dice.

Sol. Number of possible outcomes when two dice are thrown = n(S) = $6^2 = 36$

Let A be the event of getting sum as 10

Favourable outcomes are {(4,6),(5,5),(6,4)}, $n(A) = 3$

So the probability of getting sum as 10 = $\frac{3}{36} = \frac{1}{12}$

12.What is the probability of having a king and a queen when 2 cards are drawn from a pack of 52 cards

Sol. Total number of cards = 52

Drawing 2 cards from 52 can be done in 52_{c_2} ways. So n(S) = 52_{c_2}

Favourable number of cases = 4_{c_1}x4_{c_1}ways

Probability = $\frac{4_{c_1}x4_{c_1}}{52_{c_2}} = \frac{16}{1326}$

13. Three cards are drawn from a pack of 52 cards. Find the probability that

i) 3 are hearts ii) 2 clubs and 1 diamond iii) 1 heart 1 diamond and 1 club

Sol. Drawing 3 cards from a pack of 52 cards can be done in 52_{c_3} ways,

So n(S) = 52_{c_2}

There are 13 hearts, 13 spades, 13 clubs and 13 diamonds

i) Drawing 3 hearts out of 13 can be done in 13_{c_3} ways.

So, P(3 are hearts) = $\frac{13_{c_3}}{52_{c_2}}$

ii) P(2 clubs and 1 diamond) = $\frac{13_{c_2}13_{c_1}}{52_{c_2}}$

iii) P(1 heart 1 diamond and 1 club) = $\frac{13_{c_1}*13_{c_1}*13_{c_1}}{52_{c_2}}$

14. 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. Given 12 bulbs of which 5 are defective and 7 are non-defective

Drawing 3 bulbs out of 12 can be done in 12_{c_3} ways, so n(S) = 12_{c_3}

i) P(all are defective) = P(choosing all the 3 from 5) = $\frac{5_{c_3}}{12_{c_3}}$

ii) P(one is defective) = P(1 defective from 5 and 2 non defective from 7)

$$= \frac{5_{c_1}7_{c_2}}{12_{c_3}}$$

iii) P(two are defective) = P(2 from 5 defectives and 1 from 7 non defectives)

$$= \frac{7_{c_1}5_{c_2}}{12_{c_3}}$$

Exercise

1. Two bulbs are fitted in a room. Two bulbs are chosen at random from 8 bulbs having 5 good bulbs. What is the probability that the room is lit?

2. If 5 out of 30 tubes are defective and 7 of them are randomly chosen for inspection, what is the probability that one two of the defective tubes will be included?

3. What is the probability that 2 cards drawn at random from a pack of 52 cards to be one king and one queen?

4. A number is chosen from 1 to 10. Find probability that it is: a) Prime b) Multiple of 3

5. A card is drawn from a standard pack of 52 cards. Find the probability that it is i) A spade ii) A face card

6. Bag A contains 4 white and 5 black balls and Bag B contains 3 white and 4 black balls. If a ball is drawn from each bag, what is the probability that they are both of the same colour?

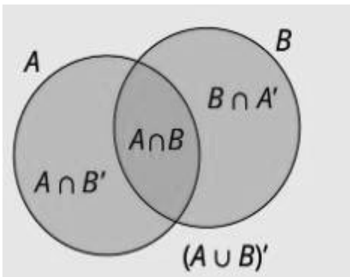
Answers

1) 25/28 2) 0.261 3) 16/1326 4) 2/5, 3/10 5) 1/4, 3/13 6) 32/63

Addition Theorem: If A and B are any two events of sample space S, then

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

Proof:



From the figure $A \cap B^1, A \cap B, B \cap A^1$ are all mutually disjoint.

$A \cup B = A \cup (B \cap A^1)$

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

$P(A \cup B) = P(A) + P(B \cap A^1)$ by Axiom iii

(Since $B \cap A^1$ and A are mutually disjoint)

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

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

Addition Theorem for 3 events: If A, B, C are the events of sample space S then

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

(Hint: Take $B \cup C$ = k, then$P(A \cup B \cup C) = P(A \cup K)$, use addition theorem for 2 events and substitute k Value again use addition theorem for 2 events for proving this theorem)

Solved Problems:

1. A card is drawn from a pack of cards. What is the probability that it is a club or an ace?

Sol. Let S represents Sample space then n(S) = 52

Let A, B represent clubs and aces respectively. $(A \cup B)$ represents getting a club or an ace.

$P(A) = \frac{13}{52}, P(B) = \frac{4}{52}, P(A \cap B) = \frac{1}{52}$

By Addition Theorem, $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

$P(A \cup B) = \frac{13}{52} + \frac{4}{52} - \frac{1}{52} = \frac{4}{13}$

2. A face recognition system fails due to Poor lighting is 0.30, Camera blur is 0.25, both together is 0.10. Find the probability that recognition fails due to either poor lighting or blur.

Sol. Let A, B represents face recognition failing due to poor lighting and camera blur respectively.

$P(A) = 0.3, P(B) = 0.25, P(A \cap B) = 0.1$

By Addition Theorem, $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

P(recognition fails due to either poor lighting or blur.) = $P(A \cup B)$

$= 0.3 + 0.25 - 0.1 = 0.45$

3. An autonomous car detects pedestrians using Camera is 0.70, Radar is 0.60, and both is 0.40. Find probability of detection by at least one sensor.

Sol. Let A, B represents an autonomous car detecting pedestrians using Camera and Radar respectively.

$P(A) = 0.7, P(B) = 0.6, P(A \cap B) = 0.4$

By Addition Theorem, $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

P(pedestrian detection by at least one sensor) = $P(A \cup B)$

$$= 0.7 + 0.6 - 0.4 = 0.9$$

4. In a survey it is found that 50% of the men like Tea, 40% like Coffee, 30% like Juice, 20% like Tea & Coffee, 15% like Coffee & Juice, 10% like Tea & Juice, 5% like all three. Find the probability that a person likes at least one drink.

Sol. Let A, B, C represent men liking Tea, Coffee, Juice respectively.

$P(A) = 0.5,\ P(B) = 0.4, P(C) = 0.3, P(A \cap B) = 0.2, P(B \cap C) = 0.15,$

$P(C \cap A) = 0.1$, $P(A \cap B \cap C) = 0.05$

From the addition theorem for 3 events, we have

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

P(person likes at least one drink) $= P(A \cup B \cup C)$

$$= 0.5 + 0.4 + 0.3 - 0.2 - 0.15 - 0.1 + 0.05 = 0.8$$

5. Three students A, B, C are in the running race. A and B have the same chance of winning and each is twice as likely to win as C. Find the probability that A or C wins.

Sol. Given $P(A) = P(B)\ and\ P(A) = 2P(C)$ and $A \cup B \cup C = S$ = sample space.

We know that $P(A) + P(B) + P(C) = 1$

$$2P(C) + 2P(C) + P(C) = 1$$

$$5P(C) = 1$$

$$P(C) = 1/5, P(A) = P(B) = 2/5$$

So, $P(A \cup C) = P(A) + P(C)$ Since A and C are mutually exclusive.

$P(A \cup C) = 2/5 + 1/5 = 3/5$

6. The Probability of horse A winning the race is 1/5 and that of horse B winning the same race is 1/4. What is the probability that i) either of them will win ii) none of them will win.

Sol. A and B are two horses. Given $P(A) = 1/5, P(B) = 1/4$

i) P(either of them will win) = $P(A \cup B)$
$= P(A) + P(B)$ (Since A and B are mutually exclusive)
$= 1/5 + 1/4 = 9/20$
ii) P(none of them will win) = 1- P(At least one or either of them will win)
$= 1 - 9/20 = 11/20$

7. An integer is chosen at random from the first 100 positive integers. What is the probability that the integer chosen is divisible by 4 or 5 ?

Sol. Let A and B represent an integer chosen that is divisible by 4 and 5 respectively.

Here $n(S) = 100$

$n(A) = \frac{100}{4} = 25, n(B) = \frac{100}{5} = 20,, n(A \cap B) = \frac{100}{20} = 5$ since LCM(4,5) = 20,

$P(A) = \frac{n(A)}{n(S)} = \frac{25}{100}$

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

$P(A \cup B) = \frac{25}{100} + \frac{20}{100} - \frac{5}{100} = \frac{40}{100} = 0.4$

8. The probability that a student passes in the Data structures examination is 2/3 and the probability that he /she will not pass in the Java examination is 5/9. The probability that he/she will pass in at least one of the examinations is 4/5. Find the probability that he /she will pass in both the examinations.

Sol. The probability of passing in Data structures P(DS) = 2/3

The probability of passing in Java P(J) = 1-5/9 = 4/9

Probability of Passing at least one of these examination = P(DS U J) = 4/5

P (DS U J) = P (DS) + P (J) – P (DS ∩ J).

4/5 = 2/3 + 4/9 - P (DS ∩ J).

P (DS ∩ J) = 2/3 + 4/9 - 4/5 = 14/45.

The probability of passing both examinations P (DS ∩ J) = 14/45

9. A committee of 5 is chosen from a group of 8 men and 4 women. What is the probability that the group contains a majority of women?

Sol. There are 12 people. Among them we have to choose 5 such that the majority would be women. Given there are 8 men and 4 women. Hence the possible selection of men and women will be either 1 man and 4 women or 2 men and 3 women.

P(1M and 4 W) = $(8_{c_1} \times 4_{c_4})/(12_{c_5})$ = 8/792

P(2M and 3 W) = $(8_{c_2} \times 4_{c_3})/(12_{c_5})$ = 112/792

Therefore the probability of selecting 5 people is 8/792 + 112/792 = 5/33

Exercise-2

1. In a network system, Probability of server failure is 0.25, Probability of database failure is 0.2 and probability of both failures is 0.1 then Find the probability that either system fails.

2. In a speech recognition system, Probability that a word is misclassified due to accent is 0.25, Probability that a word is misclassified due to background noise is 0.15 and Probability of both happening simultaneously is 0.10. What is the probability that a word is misclassified due to either accent or noise?

3. In a robot vision system, Probability that the camera detects an object is 0.6, probability that LiDAR detects the object is 0.55 and Probability both detect is 0.35 then what is the probability that the object is detected by at least one sensor?

4. In a college, 70% students play cricket, 50% play football and 30% play both. Find the probability that a student plays at least one sport.

5. A card is drawn from a standard deck of 52 cards. Find the probability that the card is a spade or a queen.

6. The probability that the contractor will get a painting contract is 0.7 and the probability that he would not get an electric contract is 0.2. The probability that he will get at least one of the contracts is 0.9 then what is the probability that he will get both the contracts?

Answers

1) 0.35 2) 0.3 3) 0.8 54) 0.9 5) 4/13 6) 0.6

Conditional Probability: If A and B are any two events of the sample space S, then happening of A after the happening of event B denoted by (B/A) and happening of B after the happening of A denoted by (A/B) are called conditional events and their corresponding probabilities are called conditional probabilities.

$$P(A/B) = \frac{P(A \cap B)}{P(B)}, P(B) > 0$$

$$P(B/A) = \frac{P(A \cap B)}{P(A)}, P(A) > 0$$

Multiplication Theorem: If A and B are any two events of sample space S, then

$$P(A \cap B) = P(A).P(B/A) \text{ or } P(B).P(A/B)$$

Independent Events: If the occurrence of one event does not influence the occurrence of the other, then the two events A and B are said to be independent of each other otherwise dependent.

Events A and B are said to be independent if $P(A \cap B) = P(A).P(B)$

Equivalent form $P(A/B) = P(A)\ and\ P(B/A) = P(B)$

Note: If A and B are independent events, then

- $A\ and\ B^C$ are independent events
- $A^C\ and\ B$ are independent events

- A^C and B^C are independent events

Ex. Throw a die twice. Let A = {max is 2} and B = {min is 2}. Are A and B independent?

Sol. First list out A and B:

A = {(1, 2), (2, 1), (2, 2)},

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

Therefore, the probabilities are

P[A] = 3/36, P[B] = 9/36, and P[A ∩ B] = P[(2, 2)] = 1/36,

Clearly, P[A ∩ B] $\neq$ P[A]P[B] and so A and B are dependent.

Note: Drawing Cards with replacement gives independent events and without replacement gives dependent events.

Solved Problems

1. If $P(A) = 3/8, P(B) = 1/3$ and $P(A \cap B) = 1/4$. Find i) $P(A^C)$ ii) $P(B^C)$ iii) $P(A^C \cup B^C)$iv) $P(A \cup B)$, v)$P(A^C \cap B^C)$ vi) $P(A/B)$ vii) $P(B/A)$ ix) $P(B^C/A)$

Sol $P(A^C) = 1 - P(A) = 1 - 3/8 = 5/8$

$P(B^C) = 1 - P(B) = 1 - 1/3 = 2/3$

$P(A^C \cup B^C) = P(A \cap B)^C = 1 - P(A \cap B) = 1 - 1/4 = 3/4$

$P(A \cup B) = P(A) + P(B) - P(A \cap B = 3/8 + 1/3 - 1/4 = 11/24$

$P(A^C \cap B^C) = P(A \cup B)^C = 1 - P(A \cup B) = 1 - 11/24 = 13/24$

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

$$P(B/A) = \frac{P(A \cap B)}{P(A)} = \frac{1/4}{3/8} = 2/3$$

$$P(B^C/A) = \frac{P(A \cap B^C)}{P(A)} = \frac{P(A) - P(A \cap B)}{P(A)} = 1 - \frac{P(A \cap B)}{P(A)} = 1 - 2/3 = 1/3$$

2. A data science team is working on a model to predict whether an email is spam or not. They are using two independent features: The presence of the word "offer" (Feature A) and the presence of a suspicious link (Feature B). The probability that an email contains the word offer" (Feature A) is 0.6. The probability that an email contains a suspicious link (Feature B) is 0.4.
   a) What is the probability that an email contains both the word "offer" and a suspicious link?
   b) What is the probability that an email contains either the word "offer" or a suspicious link or both?
   c) What is the probability that an email contains neither the word "offer" nor a suspicious link?

Sol. Given $P(A) = 0.6, P(B) = 0.4$

a) P(email contains both the word "offer" and a suspicious link) = $P(A \cap B)$

Since the events A and B are independent, $P(A \cap B) = P(A).P(B) = 0.6X0.4 = 0.24$

b) P(email contains either the word "offer" or a suspicious link or both) = $P(A \cup B)$

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0.6 + 0.4 - 0.24 = 0.76$$

c) P (email contains neither the word "offer" nor a suspicious link) = $P(A^C \cap B^C)$

$$P(A^C \cap B^C) = P(A \cup B)^C = 1 - P(A \cup B) = 1 - 0.76 = 0.24$$

3. Two balls are drawn one after the other from a bag containing 10 red, 30 white, 20 blue and 15 black balls with replacement being made after each drawing. Find the probability that i) both the balls are white ii) first ball is red and second is white iii) neither is black

Sol. Total number of balls = 10+30+20+15 = 75,

$$P(R) = 10/75, P(W) = 30/75, P(Bl) = 20/75, P(B) = 15/75$$

i) P(both the balls white) = $(w1 \cap w2) = P(w1).P(w2) = \frac{30}{75}.\frac{30}{75}$ (Since the chance for first ball to be white is $\frac{30}{75}$, again after replacement of first ball , chance for second ball to be white is also $\frac{30}{75}$

ii)P(first ball is red and second is white) = $P(R \cap W) = P(R).P(W) = \frac{10}{75} X \frac{30}{75} = \frac{4}{75}$

iii) P(neither is black) = $P(B1^C \cap B2^C) = P(B1^C) X\, P(B2^C) = \frac{60}{75} X \frac{60}{75} = \frac{16}{25}$

4. In a class, 40% read physics, 30% read maths and 10% both. If one of the students is selected at random, find the probability that i) the student read physics if it is known that he read maths ii) the student read maths if it is known that he read physics.

Sol. Let A and B be the event of students reading physics and maths respectively.

Given P(A) = 40% = 0.4, P(B) = 30% = 0.3, $P(A \cap B)$ = 10% = 0.1

i) P(the student read physics if it is known that he read maths)

$$= P(A/B) = \frac{P(A \cap B)}{P(B)} = \frac{0.1}{0.3} = 1/3$$

ii) P(the student read maths if it is known that he read physics) = $P\left(\frac{B}{A}\right) P\left(\frac{B}{A}\right)$

$$= \frac{P(A \cap B)}{P(A)} = \frac{0.1}{0.4} = 1/4$$

5. A Problem in Statistics is given to 3 students A, B and C whose chances of Solving are 1/2, 1/4, 1/3 respectively. What is the probability that the problem is solved?

Sol. Give $P(A) = 1/2, P(B) = 1/4, P(C) = 1/3$ and these events are independent events

P(the problem is Solved) = $P(A \cup B \cup C) = 1 - P(A^C \cap B^C \cap C^C)$

$$= 1 - (1/2\, X 3/4 X 2/3) = 3/4$$

Or can use Addition theorem for 3 events formula.

6. The class has 15 boys and 10 girls. These students are selected at random one after the other. Find the probability that i) First two are girls and third is a boy ii) First and third are of same gender and second is of opposite gender.

Sol. Total number of students = 15+10 = 25.

i) P(First two are girls and third is a boy) = $P(G1 \cap G2 \cap B) = \frac{10}{25}\frac{9}{24}\frac{15}{23}$

ii) P(First and third are of same gender and second is of opposite gender)

$$= P((G1 \cap B \cap G2) \cup (B1 \cap G \cap B2) = P(G1 \cap B \cap G2) + P(B1 \cap G \cap B2).$$

$$= \frac{10}{25}\frac{15}{24}\frac{9}{23} + \frac{15}{25}\frac{10}{24}\frac{14}{23}$$

7. A can hit the target once in 4 shots, B can hit 3 targets in 5 shots, C can hit two targets in 3 shots. What is the probability that 2 shots hit the target?

Sol. Give $P(A) = 1/4, P(B) = 3/5, P(C) = 2/3$

P(2 shots hit the target) = $P((A^C \cap B \cap C) \cup (B^C \cap A \cap C) \cup (C^C \cap B \cap A))$

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

$$= P(A^C).P(B).P(C) + P(B^C).P(A).P(C) + P(C^C).P(B).P(A)$$

= (3/4) .(3/5).(2/3)+(2/5).(1/4).(2/3)+(1/3).(3/5).(1/4)

$$= \frac{18+4+3}{60} \qquad = \frac{25}{60} = \frac{5}{12}$$

8. In city A, there are 40% people who have B.P., 30% have diabetes and 20% have both. A person is selected at random from the city A, i) If he/she has B.P., what is the probability that he/she has diabetes also? ii) If he/she has diabetes, then what is the probability that he/she doesn't have B.P.?

Sol. Let A, B represent B.P. and diabetes respectively.

Given $(A) = 40\% = 0.4, P(B) = 30\% = 0.3, and\ P(A \cap B) = 20\% = 0.2$

i) If he/she has B.P., what is the probability that he/she has diabetes also

$$P(B/A) = \frac{P(A \cap B)}{P(A)} = \frac{0.2}{0.4} = 1/2$$

ii) If he/she has diabetes, then what is the probability that he/she has B.P.

$$P(A/B) = \frac{P(A \cap B)}{P(B)} = \frac{0.2}{0.3} = 2/3$$

If he/she has diabetes, then what is the probability that he/she doesn't have B.P.

= 1-$P(A/B)$ = 1/3

9. Two digits are selected at random from digits 1 to 9.

i) If the sum is odd, what is the probability that 3 is one of the numbers selected?
ii) If 3 is one of the digits selected what is the probability that the sum is odd?

Sol. The given set consists of 5 odd digits {1, 3, 5, 7, 9}, 4 even digits {2, 4, 6, 8}

w.k.t. Even + odd = odd + even = odd

i) So, for getting the sum as odd, the number of cases is 5X4 = 20.

Out of these 20 cases, there are 4 cases which contains 3 as one of the numbers

i.e. {(3, 2), (3, 4), (3, 6), (3, 8)}

So, If the sum is odd, the probability that 3 is one of the numbers selected =

$$P\left(\frac{3\ is\ one\ of\ the\ numbers\ selected}{Sum\ is\ odd}\right) = \frac{4}{20} = 1/5$$

ii) If 3 is one of the digits selected we get number of cases = 8 {(3, 1), (3, 2), (3, 4)...(3,9)}

Now from these 8 cases there are 4 cases where the sum is odd.

So, if 3 is one of the digits selected what is the probability that the sum is odd =

$$P(\frac{Sum\ is\ odd}{3\ is\ one\ of\ the\ digits}) = \frac{4}{8} = 1/2$$

10. Two cards are selected at random from 12 cards numbered 1 to 12. Find the Probability that the sum is even if i) two cards are drawn together ii) two cards are drawn one after another with replacement. iii) two cards are drawn one after another without replacement.

Sol. Number of odd cards = 6 = Number of even cards

i) If two cards are drawn at a time then the number of ways of doing so is 12_{c_2} Ways.

For the sum to be even both the cards should be even or both the cards should be odd.

Choosing 2 even number cards from 6 can be done in 6_{c_2} = 15.

Similarly choosing 2 odd cards from 6 can be done in 6_{c_2} ways.

So total the number of ways of getting sum as Even is 15 + 15 = 30 ways.

And $12_{c_2} = 66.$

So, Required probability = $\frac{30}{66} = \frac{10}{22}$

ii) For the sum to be even when two cards are drawn one after the other with replacement is 6_{c_1} X 6_{c_1} + 6_{c_1} X 6_{c_1}

The number of ways in which two cards can be drawn one after the other with replacement is 12_{c_1} X 12_{c_1}

So therefore the required probability = $\frac{6_{c_1} X\, 6_{c_1} + 6_{c_1} X\, 6_{c_1}}{12_{c_1} X\, 12_{c_1}} = \frac{72}{144} = 1/2$

iii) For the sum to be even when two cards are drawn one after the other with replacement is 6_{c_1} X 5_{c_1} + 6_{c_1} X 5_{c_1}

The number of ways in which two cards can be drawn one after the other with replacement is 12_{c_1} X 11_{c_1}

So therefore the required probability = $\frac{6_{c_1} X\, 5_{c_1} + 6_{c_1} X\, 5_{c_1}}{12_{c_1} X\, 11_{c_1}} = \frac{60}{12\, X 11} = 10/22$

Exercise-3

1. The probabilities that a husband and wife will be alive 25 years from now are given by the probabilities 0.7 and 0.8 respectively. Find the probability that in 25 years i) both ii) at least one iii) none of them will be alive.

2. Find the probability of the occurrence of a number greater than two in a throw of a die, if it is known that only an odd number can occur.

3. Two dice are thrown, find the probability that the number appeared has the sum 6 if it is known that the second die always exhibits 3.

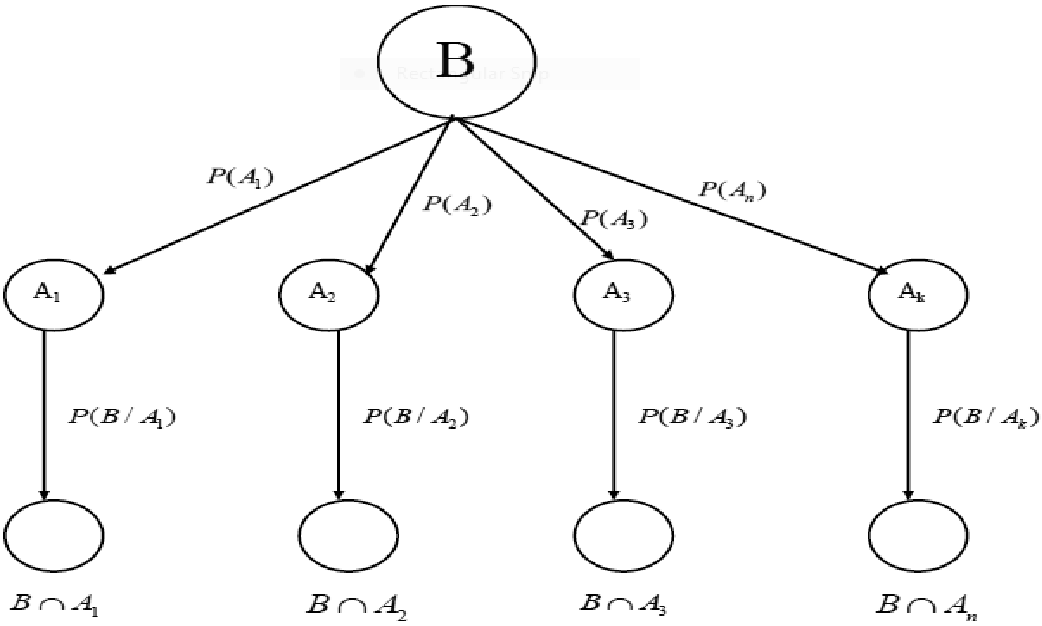
4. X and Y are independent witnesses of an event which is known to have occurred. The probabilities that they speak the truth are 0.6 and 0.8. In what percent of the cases they are likely to contradict each other.

5. A husband and a wife appear for an interview for two vacancies of the same post. The probability of husband's selection is 1/8 and that of wife's selection is 1/6 .What is the probability that i) at least one of them ii) only one of them will be selected?

Answers

1) (i) 0.56 (ii) 0.94 (iii) 0.06 2) 2/3 3) 1/6 4) 44% 5) 13/48, 1/4

Total Probability: Let $A_1, A_2, A_3, A_4 \ldots\ldots A_k$ are the mutually exclusive and exhaustive events and B$\subseteq \cup_{i=1}^{k} E_i$ then for any event B, $P(B) = P(A_1).P(B/A_1) + P(A_2).P(B/A_2) + P(A_3).P(B/A_3)\ldots\ldots.P(Ak).P(B/A_k)$



1. A person has undertaken a mining job. The probabilities of completion of the job on time with and without rain are 0.5 and 0.80 respectively. If the probability that it will rain is 0.45, then determine the probability that the mining job will be completed on time.

Sol. Let A be the event that the mining job will be completed on time and B be the event that it rains. We have,

P(B) = 0.45,

P(no rain) = P(B′) = 1 − P(B) = 1 − 0.45 = 0.55

By multiplication law of probability,

P(A|B) = 0.5, P(A|B′) = 0.80

Since, events B and B′ form partitions of the sample space S, by total probability theorem, we have P(A) = P(B) P(A|B) + P(B′) P(A|B′)

= 0.45 × 0.5 + 0.55 × 0.8

= 0.225 + 0.440 = 0.665

So, the probability that the job will be completed on time is 0.665.

2. A factory has 3 machines, Machine A_1 produces 30% items, defect rate is 3%, Machine A_2 produces 50% items, defect rate is 2%, Machine A_3 produces 20% items, and defect rate is 1%. Find probability that a randomly chosen item is defective.

Sol. $P(D) = P(D|A_1)P(A_1)+P(D|A_2)P(A_2)+P(D|A_3)P(A_3)$

P(D) = (0.03) (0.3) + (0.02) (0.5) + (0.01) (0.2) = 0.021

Bayes theorem:

Let $A_1, A_2, A_3, A_4 \ldots\ldots A_k$ are the mutually exclusive and exhaustive events of sample space S and B is any arbitrary event such that $B \subseteq \cup_{i=1}^{k} A_i$ and $P(B) > 0$, then the probability of A_i for given B is

$$P(A_i/B) = \frac{P(A_i).P(B/A_i)}{\sum_{i=1}^{k} P(A_i).P(B/A_i)}$$

Bayes Theorem is also known as the Rule of inverse probability. Here $P(A_i)$ is a priori probability, $P(B/A_i)$ likelihoods and $P(A_i/B)$ posterior probabilities

1. Given that P(Rain) = 0.3, P(Cloudy | Rain) = 0.75, P(Cloudy | No Rain) = 0.25 If it is cloudy, what is the probability it rains?

Sol. Given that P(Rain) = 0.3, P(Cloudy | Rain) = 0.75, P(Cloudy | No Rain) = 0.25

Then the probability(it rains if it is cloudy)

$$= \frac{P(Rain).PCloudy/Rain)}{P(Rain).P(cloudy/Rain)+P(No\ rain).P(cloudy/No\ ain)}$$

$$= \frac{0.3\ X\ 0.75}{0.3X0.75 + 0.7X0.25} = 0.5625$$

2. In a neighbourhood, 85% children were falling sick due to flu and 15% due to measles and no other disease. The probability of observing rashes for measles is 0.95 and for flues 0.1. If a child develops rashes, find the child's probability of having flu.

Sol. Let, F: children with flu M: children with measles,

R: children showing the symptom of rash

P(F) = 85% = 0.85, P(M) = 15% = 0.15, P(R|F) = 0.1, P(R|M) = 0.95

P(Child having flu if he/she develops rashes) = $P(F/R) = \frac{P(F).P(R/F)}{P(F).P(R/F)+P(M).P(R/M)}$

$$P\left(\frac{F}{R}\right) = \frac{0.85X\ 0.1}{0.85X0.1\ +0.15X0.95} = 0.3936$$

3. Of the three candidates, the chances that a banker, an economist, and a chartered accountant will be appointed as the finance head are 0.5, 0.3, and 0.2 respectively. The probabilities that financial performance improves under them are 0.3, 0.7, and 0.8 respectively.

   i) Determine the probability that financial performance improves.

   ii) If financial performance improves, what is the probability that the finance head is a chartered accountant?

Sol. Let A, B, C be the events that a banker, an economist and a chartered accountant will be appointed as the finance head. Then $P(A) = 0.5,\ P(B) = 0.3,\ P(C) = 0.2$

Let F represents improvement in financial performance.

The probability that the financial performance improves if they are appointed as finance head are $P(F/A) = 0.3, P(F/B) = 0.7, P(F/C) = 0.8$

i) P(the financial performance improves) = $P(F)$

$$= P(A).P\left(\frac{F}{A}\right) + P(B).P\left(\frac{F}{B}\right) + P(C).P\left(\frac{F}{C}\right)$$

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

= 0.52

ii) The probability that if financial performance improves, it is by the chartered accountant

$$= P\left(\frac{C}{F}\right) = \frac{P(C).P\left(\frac{F}{C}\right)}{P(A).P\left(\frac{F}{A}\right)+P(B).P\left(\frac{F}{B}\right)+P(C).P\left(\frac{F}{C}\right)}$$

$$= \frac{(0.2)(0.8)}{0.52} = 0.3076$$

4. The probability that Doctor X will diagnose disease A correctly is 70%. The probability that a patient will die by his/her treatment after a correct diagnosis is 30% and the probability of death by wrong diagnosis is 60%. A patient of doctor X who had disease A died, then what is the probability that his disease was diagnosed correctly.

Sol. Let $E1$and $E2$ be the event that disease A was diagnosed correctly and not correctly respectively. Let B be the event that the patient of Doctor X died.

$P(E1) = 0.7, P(E2) = 0.3$

$$P(B/E1) = 0.3, P(B/E2) = 0.6$$

Now by Bayes theorem, required probability is given by

$$P(E1/B) = \frac{P(E1).P(B/E1)}{P(E1).P(B/E1)+P(E2).P(B/E2)}$$

$$P\left(\frac{E1}{B}\right) = \frac{(0.7)(0.3)}{(0.7)(0.3)+(0.3)(0.6)} = 0.5384$$

5. The results of an investigation by an expert on a fire accident in a restaurant are summarized below:

(i) P(There should have been short circuit) = 0.7
(ii) P(LPG cylinder explosion) = 0.3
(iii) Chance of fire accident is 30% given a short circuit and 95% given an LPG explosion.
Based on these, what do you think are the most probable causes of fire? Statistically justify your answer.

Sol. Let E1, E2 represent short circuit and LPG explosions respectively. A represents a fire accident.

Then $P(E1) = 0.7, P(E2) = 0.3, P(A/E1) = 0.3, P(A/E2) = 0.95$

Then by Bayes theorem, required probabilities are given by $P(E1/A)$ and $P(E2/A)$

where $P(E1/A) = \frac{P(E1).P(A/E1)}{P(E1).P(A/E1)+P(E2).P(A/E2)} = \frac{(0.7)(0.3)}{(0.7)(0.3)+(0.3)(0.95)} = 0.424$

$$P(E2/A) = \frac{P(E2).P(A/E2)}{P(E1).P(A/E1)+P(E2).P(A/E2)} = \frac{(0.3)(0.95)}{(0.7)(0.3)+(0.3)(0.95)} = 0.576$$

Conclusion: The most probable cause of the fire is an LPG explosion (≈ 57.6%) rather than a short circuit (≈ 42.4%),

6. In a soap factory Machines X, Y, Z manufactures 30%, 30% and 40% of soaps. Of their total outputs. 5%, 4%, 3% are defective. A soap is drawn at random and found to be defective. What is the probability that it is manufactured by i) Machine X, ii) Machine Z.

Sol. Let E1, E2, E3 represent soaps manufactured by machine X, Y, Z respectively.

Let D represents defective.

From the given $P(E1) = 0.3, P(E2) = 0.3, P(E3) = 0.4,\ P\left(\frac{D}{E1}\right) = 0.05,$
$P(D/E2) = 0.04, P(D/E3) = 0.03$
Then by Bayes theorem, required probabilities are given by $P(E1/D)\ and\ P(E3/D)$ where

$$P\left(\frac{E1}{D}\right) = \frac{P(E1).P\left(\frac{D}{E1}\right)}{P(E1).P\left(\frac{D}{E1}\right) + P(E2).P\left(\frac{D}{E2}\right) + P(E3).P\left(\frac{D}{E3}\right)}$$

$$= \frac{(0.3)(0.05)}{(0.3)(0.05)+(0.3)(0.04)+(0.4)(0.03)} = 0.3846$$

$$P(\frac{E3}{D}) = \frac{P(E3).P(D/E3)}{P(E1).P(D/E1) + P(E2).P(D/E2) + P(E3).P(D/E3)}$$

$$= \frac{(0.4)(0.03)}{(0.3)(0.05) + (0.3)(0.04) + (0.4)(0.03)} = 0.307$$

Exercise

1. In a certain college, 5% of the boys and 2% of the girls are taller than 1.7m, furthermore 60% of the students are girls. Now if a student is selected at random and is taller than 1.7m, then what is the probability that the student is a girl?

2. Companies A, B, C produce 30%, 40% and 30% of the cars respectively. It is known that 2%, 3% and 4% of the cars produced from A, B, C are defective. i) What is the probability that a car purchased is defective? ii) If a car purchased is found to be defective, what is the probability that this car is produced by company C?

3. A box I contains 2 black and 3 white marbles and a box II contains 4 black and 5 White marbles. One marble is drawn at random from one of the boxes and it is found to be white. Find the probability that the white marble drawn is from box II.

4. A businessman goes to hotels A, B, C 30%, 40%, 30% of the time respectively. It is known that 4%, 3%, 5% of the rooms in A, B, C hotels have faulty plumbings. What is the probability that a business man's room having faulty plumbing is assigned to hotel A?

5. If 5% and 8% of the men and women in a City A are colour blind. A colour blind is selected at random. What is the probability that the person is a female? Assume that male and females are in equal numbers.

Answers

1) 3/8 2) 25/52 3) 0.03, 0.4 4) 4/13 5) 8/13

Unit-II

Random Variables

Random Variables:

A Random Variable means a real number x which is connected with the outcome of a random experiment.

Let S be the sample space associated with random experiment, the rule by which for each outcome of random experiment a real number is assigned is known as random variable.

Definition: A random variable (r.v.) is a function $X(x)$ with domain S and range $(-\infty, \infty)$ such that for every real number a, the event $(x, X(x) \leq a) \in S$.

Example 1: Consider a random experiment *consisting* of tossing a coin twice. Let random variable X denote the number of times a tail appears. The sample space $S = \{HH, HT, TH, TT\}$. Find the values of random variables and its probabilities.

Sol. Sample Space $S = \{HH, HT, TH, TT\}$, Here the random variable takes the values 0,1 and 2.

By definition of random variable, $X(HH) = 0, X(HT) = 1, X(TH) = 1,\ X(TT) = 2.$

$$P(X = 0) = \frac{1}{4}, \quad P(X = 1) = \frac{2}{4}, \quad and\ P(X = 1) = \frac{1}{4}$$

The probability of random variables can be shown in tabular form as

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

Example 2: Consider a random experiment *consisting* of throwing a pair of dice. Let random variable X denote the sum of the numbers appeared on dice. Find the values of random variables and its probabilities.

Sol. The sample space $S = \{(1,1),(1,2),(1,3),(1,4),(1,5),(1,6),(2,1),(2,2),(2,3),(2,4),(2,5),(2,6), \ldots\ldots(6,1),(6,2),(6,3),(6,4),(6,5),(6,6),\}$

Here the random variable is the sum of the numbers appeared on dice and it takes the values 2, 3, 4…12

By definition of random variable, $X\big((1,1)\big) = 2, X\big((1,2),(2,1)\big) = 3, \ldots X\big((6,6)\big) = 12$

$$P(X = 2) = \frac{1}{36}, \quad P(X = 3) = \frac{2}{36}, \quad and\; P(X = 6) = \frac{1}{36}$$

The probability of random variables can be shown in tabular form as

x	2	3	4	5	6	7	8	9	10	11	12
$P(X = x)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

Note: If X and Y are two random variables on the Sample Space S, K is any constant,

then $X+Y, X+K, KX$ and $KY$are also random variables on S and are defined as

$(X+Y)(x) = X(x)+Y(x),$

$(X+K)(x) = X(x)+K, (KX)(x) = KX(x)$ and $(KY)(x) = KY(x)$.

Types of Random Variables

Random variables can be classified into two types. They are Discrete random variable and Continuous random variable.

1. **Discrete random variable:** A random variable X which can take only a finite number of discrete values in an interval is called a Discrete random variable.

Example 1: In tossing two coins, define a random variable X which counts the number of heads. The Sample Space $S = \{HH, HT, TH, TT\}$ here the random variable will take the values $X = \{0, 1, 2\}$

Example 2: In throwing two dice, define a random variable Z which is the sum of the numbers on faces of the dice. The sample space

$$S = \{(1,1), (1,2), (1,3), (1,4), (1,5), (1,6), (2,1), (2,2), (2,3), (2,4), (2,5), (2,6), \ldots..(6,1), (6,2), (6,3), (6,4), (6,5), (6,6)\}$$

here the random variable will take the values

$$Z = \{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$$

2. **Continuous random variable:** A random variable X which can take values continuously, all possible values in a given interval is called a Continuous random variable.

Example 1: The age/weight of a randomly selected student from a class.

Example 2: The temperature at a randomly selected place.

Probability Distribution of discrete random variable:

Suppose X is a discrete random variable, the probability distribution of X is set of all possible values of X along with their respective probabilities.

Example 1 The probability distribution is given by

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

Example 2, the probability distribution is given by

x	2	3	4	5	6	7	8	9	10	11	12
$P(X = x)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

Probability Function of Discrete random variable:

A probability function of a discrete random variable is called the Probability Mass Function (PMF).

Definition: Let X be a discrete random variable. The probability function of X is called probability mass function (PMF) is denoted by $p(x)$and is defined as: $p(x) = P(X = x)$. It gives the probability that the random variable X takes a specific value x.

Properties of PMF

1. Non-negative: $p(x) \geq 0$ for all x.
2. Sum equals 1: $\sum p(x) = 1$
3. Defined only for discrete values: x takes countable values (e.g., 0, 1, 2, …)

Example 1: In games like dice, cards, outcomes are **discrete**.

PMF helps calculate chances of winning/losing and expected gains or losses.

Example 2: Let X = number of defective items in a batch.

PMF helps to estimate defect probabilities and to improve production processes. Used in Six Sigma and industrial statistics.

Distribution Function of a Discrete Random Variable:

The distribution function of a discrete random variable is called the Cumulative Distribution Function (CDF). Let X be a discrete random variable. The distribution function is denoted by $F(x)$ and is defined as

$F(x) = P(X \leq x)$. It gives the probability that X takes a value less than or equal to x.

Properties of CDF

1. Non-decreasing function: $\boldsymbol{F(x_1) \leq F(x_2), if\ x_1 < x_2}$
2. Limits: $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to \infty} F(x) = 1$

Measure of a discrete random variable

The measures of a discrete random variable are numerical quantities that describe its behavior—such as its average value, spread, and shape.

1. Mean (Expected value) of a random variable is average outcome. It is denoted by μ *or* $E(X)$ and is defined as $E(X) = \sum xp(x)$ or $E(X) = \sum xP(X = x)$.

Interpretation: Long-run average value of the discrete random variable.

2. Variance measures the spread or dispersion of values around the mean. It is denoted by σ^2 *or* $V(X)$ and is defined as $V(X) = \sum (x - \mu)^2 p(x)$ or $V(X) = E(X^2) - \left(E(X)\right)^2$. where $\mu = E(X)$.

Interpretation: How far values deviate from the mean.

Other measures are Standard deviation (measures spread in same units), Skewness (measures asymmetry of distribution) and Kurtosis (measures peakedness or flatness).

Solved Problems

1. A person tosses a fair coin 3 times. Let X denote the number of heads obtained. Find the probability distribution of X. Also find $P(X = 2)$ and $P(1 \leq X < 3)$.

Sol. Here X denote the number of heads obtained when a fair coin tossed 3 times. Then X takes the value $X = \{0, 1, 2, 3\}$. The sample space

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

For $X = 0$, the favorable item is TTT. Then $P(X = 0) = \frac{1}{8}$

similarly for other values of random variable X are given by

$P(X = 1) = \frac{3}{8}$, $P(X = 2) = \frac{3}{8}$ and $P(X = 3) = \frac{1}{8}$

The probability distribution is given by

x	0	1	2	3
$P(X = x)$	1	3	3	1

$P(X = 2) = \frac{3}{8}$ and $P(1 \leq X < 3) = P(X = 1) + (X = 2) = \frac{3}{8} + \frac{3}{8} = \frac{6}{8} = \frac{3}{4}$

2. A call center receives calls with probability function $P(X = x) = \frac{x}{10}$, $x = 0,1,2,3,4$. Verify whether the probability function is valid. Compute mean and variance if probability function is valid.

Sol. Here X is the random variable which takes the values 0,1,2,3,4.

The probability function is given by

$P(X = 0) = \frac{0}{10} = 0, P(X = 1) = \frac{1}{10}, P(X = 2) = \frac{2}{10}$

$P(X = 3) = \frac{3}{10}$ and $P(X = 4) = \frac{4}{10}$

And $P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4)$

$$= 0 + \frac{1}{10} + \frac{2}{10} + \frac{3}{10} + \frac{4}{10} = 1$$

Since sum of all probabilities of a random variable is equal to 1. The given probability function is valid.

The probability distribution is given by

x	0	1	2	3	4
$P(X = x)$	0	$\frac{1}{10}$	$\frac{2}{10}$	$\frac{3}{10}$	$\frac{4}{10}$

The mean of a random variable X is given $\mu = E(X) = \sum xP(X = x)$

$$\mu = E(X) = 0\left(\frac{0}{10}\right) + 1\left(\frac{1}{10}\right) + 2\left(\frac{2}{10}\right) + 3\left(\frac{3}{10}\right) + 4\left(\frac{4}{10}\right) = 3$$

The variance of a random variable X is given $V(X) = \sum(x - \mu)^2 * p(x)$ *or*

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

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

$$= 0^2\left(\frac{0}{10}\right) + 1^2\left(\frac{1}{10}\right) + 2^2\left(\frac{2}{10}\right) + 3^2\left(\frac{3}{10}\right) + 4^2\left(\frac{4}{10}\right) - 3^2$$

$$= 1$$

3. An online store records daily order:

X	1	2	3	4
P(X = x)	0.2	0.3	0.3	0.2

Find mean number of daily orders and $P(X > 3)$

Sol. The mean of a random variable X is given $\mu = E(X) = \sum xP(X = x)$

$\mu = E(X) = 1(0.2) + 2(0.3) + 3(0.3) + 4(0.2) = 2.5$ and $P(X > 3) = P(X = 4) = 0.2$

4. Find $P(X < 2), P(X \geq 2)$and $F(2)$ of the random variable for the probability distribution

x	0	1	2	3
$P(X = x)$	0	k	$2k$	$3k$

Sol. Since sum of all the probabilities of a random variable is one

$$P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) = 0 + k + 2k + 3k = 1 \Rightarrow 6k = 1 \Rightarrow k = \frac{1}{6}$$

$$P(X < 2) = P(X = 0) + P(X = 1) = 0 + k = k = \frac{1}{6}$$

$$P(X \geq 2) = P(X = 2) + P(X = 3) = 2k + 3k = 5k = \frac{5}{6}$$

$$F(2) = P(X \leq 2) = P(X = 0) + P(X = 1) + P(X = 2)$$

$$= 0 + k + 2k = 3k = \frac{3}{6} = 0.5$$

5. A shop records number of customers per hour:

x	0	1	2	3
$P(X = x)$	0.1	0.3	0.4	0.2

Find the probability that at most two customers arrive per hour.

Sol. Probability that at most two customers arrive per hour is given by cumulative distributive function $F(2) = P(X \leq 2) = P(X = 0) + P(X = 1) + P(X = 2)$

$$= 0.1 + 0.3 + 0.4 = 0.8$$

Probability that at most two customers arrive per hour is 80%.

Mean and variance of linear combination of Random variables

Theorem: If a and b are any constants, then

$$E(aX + b) = aE(X) + b$$

$$Var(aX + b) = a^2 \operatorname{var}(X)$$

1. If X any random variable with mean 5 and Variance 2, calculate the mean and Variance of $Y = 2X + 3$.

Sol. $E(X) = 5\ ;\ \sigma^2(X) = 2$

$E(Y) = E(2X+3) = 2E(X)+3 = 2\times5+3 = 13$

$\sigma^2(Y) = \sigma^2(2X+3) = 2^2\,\mathrm{var}(X) = 4\times2 = 8$

Exercise

1. A random variable X has the following probability distribution

X	0	1	2	3
P(X	0.2	0.4	0.3	0.1

Find the mean of X and $P(X \geq 2)$.

2. A fair coin is tossed 4 times. Let X denote the number of tails obtained.

Find $P(X = 3)$ and $P(X < 2)$.

3. The probability distribution of a random variable X is given below:

X	1	2	3	4
P(X = x)	k	2k	3k	4k

Find the value of k and $P(X \leq 3)$.

4. A delivery service records the number of parcels delivered per hour:

X	0	1	2	3	4
P(X = x)	0.1	0.2	0.3	0.2	0.2

Find the variance of X.

5. The probability distribution of a random variable X is given by:

X	0	1	2	3
P(X = x)	0.25	0.35	0.25	0.15

Find $F(2)$ and $P(X > 1)$.

Answers: -

1) Mean = 1.3, $P(X \geq 2) = 0.4$, 2) $P(X = 3) = \frac{1}{4}$, $P(X < 2) = \frac{5}{16}$,

3) $k = \frac{1}{10}$, $P(X \leq 3) = \frac{6}{10} = 0.6$, 4) 1.56, 5) $F(2) = 0.85$, $P(X > 1) = 0.40$

Probability Function of Continuous random variable:

A probability function of a continuous random variable is called the Probability density Function (PDF).

Definition: Consider the small interval $\left[x - \frac{dx}{2}, x + \frac{dx}{2}\right]$ of length dx round the point x, let $f(x)$ be any continuous function of x so that $f(x)dx$ represents the probability that the variable X falls in the infinitesimal interval $\left[x - \frac{dx}{2}, x + \frac{dx}{2}\right]$.

It can be represented as $P\left(x - \frac{dx}{2} \leq X \leq x + \frac{dx}{2}\right) = f(x)dx$, then $f(x)$ is called Probability density Function (PDF).

Properties of PDF

1. Non-negative: $f(x) \geq 0$ for all x.
2. Sum equals 1: $\int_{-\infty}^{\infty} f(x)dx = 1$

Distribution Function of a Continuous Random Variable:

The distribution function of a continuous random variable is called the Cumulative Distribution Function (CDF). Let X be a continuous random variable. The distribution function is denoted by $F(x)$ and is defined as $F(x) = P(X \leq x) = \int_{-\infty}^{x} f(u)du$. and $f(x) = \frac{dF(x)}{dx}$

It gives the probability that X takes a value less than or equal to x.

Properties of CDF

1. Non-decreasing function: $F(x_1) \leq F(x_2), if\ x_1 < x_2$
2. Limits: $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to \infty} F(x) = 1$
3. Probability between two values: $P(a < X \leq b) = F(b) - F(a)$
4. $F(x)$ is continuous for continuous random variables.

Measure of a continuous random variable

The measures of a continuous random variable **are** numerical quantities that describe its behavior—such as its average value, spread, and shape.

1. **Mean (Expected value)** of a random variable is average outcome. It is denoted by $\boldsymbol{\mu\ or\ E(X)}$ and is defined as $E(X) = \int_{-\infty}^{\infty} xf(x)dx$.

Interpretation: Long-run average value of the continuous random variable.

2. **Variance** measures the spread or dispersion of values around the mean. It is denoted by σ^2 or V(X) and is defined as $V(X) = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx$ or

$V(X) = \int_{-\infty}^{\infty} x^2 f(x)dx - \left(\int_{-\infty}^{\infty} xf(x)dx\right)^2$. where $\mu = E(X)$.

Interpretation: How far values deviate from the mean.

For Continuous random variables also, the other measures are Standard deviation (measures spread in same units), skewness (measures asymmetry of distribution) and Kurtosis (measures peakedness or flatness).

Solved Problems

1. A continuous random variable X has probability density function $f(x) = \begin{cases} kx, & 0 \leq x \leq 2 \\ 0, & Otherwise \end{cases}$

Find the value of k and $P(1 \leq X \leq 2)$.

Sol. For a continuous random variable, we have $\int_{-\infty}^{\infty} f(x)dx = 1$

$$\Rightarrow \int_{-\infty}^{0} f(x)dx + \int_{0}^{2} f(x)dx + \int_{2}^{\infty} f(x)dx = 1 \Rightarrow 0 + \int_{0}^{2} f(x)dx + 0 = 1$$

$$\Rightarrow \int_{0}^{2} kxdx = 1$$

$$k\left[\frac{x^2}{2}\right]_0^2 = 1 \Rightarrow k \cdot \frac{4}{2} = 1 \Rightarrow 2k = 1 \Rightarrow k = \frac{1}{2}$$

By definition $P(1 \le X \le 2) = \int_1^2 f(x)dx = \int_1^2 \frac{x}{2}dx = \frac{1}{2}\left[\frac{x^2}{2}\right]_1^2 = \frac{1}{2}\left(\frac{4}{2}-\frac{1}{2}\right) = \frac{1}{2}\left(\frac{3}{2}\right) = \frac{3}{4}$

2. The time (in minutes) a person waits for a bus is a continuous random variable with PDF $f(x) = \frac{1}{10}e^{\frac{-x}{10}}, x \ge 0.$ Find the probability that the person waits less than 5 minutes.

Sol. The probability that the person waits less than 5 minutes is given by

$\int_0^5 f(x)\, dx = \int_0^5 \frac{1}{10} e^{\frac{-x}{10}} dx$

$\mathrm{P(X < 5)} = \int_0^5 \left(\frac{1}{10}\right) \mathrm{e}^{-\frac{x}{10}} \mathrm{dx} = 1 - \mathrm{e}^{-0.5} \approx 0.3935$

3. Let the PDF of a continuous random variable X be $f(x) = \begin{cases} 2x, & 0 \le x \le 1 \\ 0, & Otherwise \end{cases}$ Find the distribution function.

Sol. $F(x) = P(X \le x) = \int_{-\infty}^{x} f(t)dt$

For $x < 0$: $F(x) = 0$

For $0 \le x \le 1$: $F(x) = \int_0^x 2tdt = x^2$

For $x > 1$: $F(x) = 1$

The distribution function is given by $F(x) = \begin{cases} 0, & x < 0 \\ x^2, & 0 \le x \le 1 \\ 1, & x > 1 \end{cases}$

4. Find the probability density function for the given cumulative distribution function of a continuous random variable $F(x) = \begin{cases} 0, & x < 0 \\ 1 - e^{-2x}, & x \ge 0 \end{cases}$

Sol. By definition $f(x) = \frac{dF(x)}{dx} \Rightarrow f(x) = \frac{d\left(1-e^{-2x}\right)}{dx} = 2e^{-2x}$

$$f(x) = \begin{cases} 0, & x < 0 \\ 2e^{-2x}, & x \geq 0 \end{cases}$$

5. A continuous random variable X has PDF $f(x) = \begin{cases} kx(2-x), & 0 \leq x \leq 2 \\ 0, & Otherwise \end{cases}$

Find $k, F(x), Mean\ and\ Variance\ of\ X.$

Sol. For a continuous random variable, we have $\int_{-\infty}^{\infty} f(x)dx = 1$

$$\Rightarrow \int_{-\infty}^{2} f(x)dx + \int_{0}^{2} f(x)dx + \int_{2}^{\infty} f(x)dx = 1 \Rightarrow 0 + \int_{0}^{2} f(x)dx + 0 = 1$$

$$\Rightarrow \int_{0}^{2} kx(2-x)dx = 1$$

$$\int_{0}^{2} kx(2-x)dx = 1 \Rightarrow k\int_{0}^{2} (2x - x^2)dx = 1 \Rightarrow k\left(2\frac{x^2}{2} - \frac{x^3}{3}\right) = 1$$

$$\Rightarrow k\left(4 - \frac{8}{3}\right) = 1 \Rightarrow k = \frac{3}{4}$$

Distribution Function $F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)dt = \int_{-\infty}^{0} f(t)dt + \int_{0}^{x} f(t)dt$

$$F(x) = \int_{0}^{x} f(t)dt = \int_{0}^{x} \frac{3}{4}(2x - x^2)dt = \frac{3}{4}\left(x^2 - \frac{x^3}{3}\right) \quad 0 \leq x \leq 2$$

Expected value or mean of a random variable is given by

$$E(X) = \int_{-\infty}^{\infty} xf(x)dx = \int_{0}^{2} \frac{3}{4}x(2x - x^2)dx = \int_{0}^{2} \frac{3}{4}(2x^2 - x^3)dx$$

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

Variance of a random variable is given by

$$V(X) = \int_{-\infty}^{\infty} x^2 f(x)dx - \left(\int_{-\infty}^{\infty} xf(x)dx\right)^2 = \int_0^2 \frac{3}{4}x^2(2x - x^2)dx - 1$$

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

$$V(X) = = \int_0^2 \frac{3}{4}(2x^3 - x^4)dx - 1 = \frac{3}{4}\left(2\frac{x^4}{4} - \frac{x^5}{5}\right) - 1 = \frac{3}{4}\left(2\frac{2^4}{4} - \frac{2^5}{5}\right) - 1$$

$$= \frac{1}{5}$$

Joint Distribution (Discrete rv 's)

- Let X and Y be two discrete rv's defined on the sample space of an experiment.
- The **joint probability mass function** $p(x, y)$ is defined for each pair of numbers (x, y) by

$p(x, y)$ = P(X = x and Y = y) such that

$p(x, y) \geq 0$ and $\sum_x \sum_y p(x, y) = 1$

- The marginal probability mass function of X denoted by $p_X(x)$ and is given by

  $p_X(x) = \sum_y p(x, y)$ for each possible value of x.

- The marginal probability mass function of Y denoted by $p_Y(y)$ and is given by

  $p_Y(y) = \sum_x p(x, y)$ for each possible value of y

- Two discrete random variables X and Y are said to be **independent** if for every pair of x and y values $p(x, y) = p_X(x) \cdot p_Y(y)$
- If it is not satisfied, then X and Y are said to be dependent.

Conditional Distributions

- Let X and Y be two discrete rv's with joint pmf $p(x, y)$ and marginal probability distribution of X is $p_X(x)$.

- Then for any X value of x for which $p_X(x) > 0$, the conditional probability mass function of Y when X = x is

$$p_{Y|X}(y|x) = \frac{p(x, y)}{p_X(x)}$$

Joint Probability Distribution Function

Let (X, Y) be a two-dimensional discrete random variable then their joint distribution function is given by

$$F_{XY}(x, y) = P(X \le x, Y \le y) = \sum_{X\le x}\sum_{Y\le y} p(x, y)$$

Marginal Distribution Function

$$F_X(x) = \sum_{X\le x, \forall y} P(X \le x, Y = y)$$

$$F_Y(y) = \sum_{Y\le y, \forall x} P(X = x, Y \le y)$$

1. Let $x_1$and x_2 have the joint probability distribution f(x_1 , x_2)

		x_1		
		0	1	2
x_2	0	0.1	0.4	0.1
	1	0.2	0.2	0

a) Find P(x_1+ x_2> 1) b) Find the Probability distribution $f_1(x_1) = P(\mathrm{X} = x_1)$ of the individual random number x_1. c) Find the Conditional Probability distribution of x_1 given x_2 = 1. Are x_1 and x_2 independent?

Sol. a) The event x_1+ x_2>1 is composed of the pairs of values (1, 1), (2, 0) and (2, 1). Adding their Corresponding probabilities

$P(x_1 + x_2 > 1) = f(1, 1) + f(2, 0) + f(2, 1) = 0.2 + 0.1 + 0 = 0.3$

b) Since the event $x_1 = 0$ is composed of two pairs of values (0, 0) and (0, 1) , we add their corresponding probabilities to obtain

$P(x_1 = 0) = f(0, 0) + f(0, 1) = 0.1 + 0.2 = 0.3$

Continuing, We obtain $P(x_1 = 1) = 0.6$ and $P(x_1 = 2) = 0.1$

i.e. $f_1(0) = 0.3$, $f_1(1) = 0.6$, $f_1(2) = 0.1$ is the probability distribution of x_1.

x_1	0	1	2
$P(X = x_1)$	0.3	0.6	0.1

c) Joint Probability distribution $f(x_1, x_2)$ of x_1 and x_2 with Marginal Distributions

$f_1(x_1)$ is Marginal Distributions of x_1 and $f_2(x_2)$ is Marginal Distributions of x_2

		x_1			**Total**
		0	1	2	**$f_2(x_2)$**
x_2	0	0.1	0.4	0.1	0.6
	1	0.2	0.2	0	0.4
Total $f_1(x_1)$		0.3	0.6	0.1	1.0

$$f_1(0 \mid 1) = \frac{f(0,1)}{f_2(1)} = \frac{0.2}{0.4} = 0.5$$

$$f_1(1 \mid 1) = \frac{f(1,1)}{f_2(1)} = \frac{0.2}{0.4} = 0.5$$

$$f_1(2 \mid 1) = \frac{f(2,1)}{f_2(1)} = \frac{0}{0.4} = 0$$

Since $f_1(0 \mid 1) = 0.5 \neq 0.3 = f_1(0)$

The Conditional Probability distribution is not free of the value x_2

Also $f(0,1) = 0.2 \neq (0.3)(0.4) = f_1(0)f_2(1)$

Therefore, the two random variables x_1 and x_2 are dependent.

2. The joint probability mass function (PMF) of two discrete random variables X and Y is defined as:

$$P(X = x, Y = y) = \begin{cases} c(x+y), & \text{for } x = 1,2 \text{ and } y = 1,2 \\ 0, & \text{otherwise} \end{cases}$$

a) Find the value of the constant c.

b) Find the marginal distributions $P_X(x)$ and $P_Y(y)$.

c) Find $P(X = 1, Y = 2)$ and $P(X = 2)$.

Sol. a) P(1,1) = c(1+1) = 2c

P(1,2) = c(1+2) = 3cP(1,2) = c(1+2) = 3cP(1,2) = c(1+2) = 3c

P(2,1) = c(2+1) = 3cP(2,1) = c(2+1) = 3cP(2,1) = c(2+1) = 3c

P(2,2) = c(2+2) = 4cP(2,2) = c(2+2) = 4cP(2,2) = c(2+2) = 4c

$$\sum_{x=1}^{2}\sum_{y=1}^{2} P(X = x, Y = y) = 1$$

Which gives $c = 1/12$

b) Marginal distribution of X

$$P_X(1) = P(1,1) + P(1,2)$$
$$= \frac{1}{12}(2) + \frac{1}{12}(3) = \frac{5}{12}$$

$$P_X(2) = P(2,1) + P(2,2)$$
$$= \frac{1}{12}(3) + \frac{1}{12}(4) = \frac{7}{12}$$

Therefore,

x	1	2
$P_X(x)$	$\frac{5}{12}$	$\frac{7}{12}$

Marginal distribution of Y

$$P_Y(1) = P(1,1) + P(2,1)$$
$$= \frac{1}{12}(2) + \frac{1}{12}(3) = \frac{5}{12}$$

$$P_Y(2) = P(1,2) + P(2,2)$$
$$= \frac{1}{12}(3) + \frac{1}{12}(4) = \frac{7}{12}$$

Therefore,

y	1	2
$P_Y(y)$	$\frac{5}{12}$	$\frac{7}{12}$

c) Find $P(X = 1, Y = 2)$ and $P(X = 2)$

$$P(X = 1, Y = 2) = c(1+2) = \frac{1}{12}(3) = \frac{1}{4}$$

Also, $P(X = 2) = \frac{7}{12}$

Exercise

1. A large insurance agency services a number of customers who have purchased both a homeowner's policy and an automobile policy from the agency. For each type of policy, a deductible amount must be specified. For an automobile policy, the choices are \$100 and \$250, whereas for a homeowner's policy, the choices are 0, \$100, and \$200. Suppose an individual with both types of policy is selected at random from the agency's files. Let X the deductible amount on the auto policy and Y the deductible amount on the homeowner's policy Suppose the joint pmf is given by

p(x,y)		Y		
		0	100	200
X	100	0.20	0.10	0.20
	250	0.05	0.15	0.30

Find (i) Marginal probabilities of X and Y.
(ii) P($Y \geq 100$)
(iii) Are X and Y independent ?
(iv) Conditional distribution of Y = 100 for X = x.

2. The joint probability distribution of two random variables X and Y is given by
P(X = 0,Y = 1) = 1/3, P(X = 1,Y = -1) = 1/3, P(X = 1,Y = 1) = 1/3.

Find i) Marginal distribution of X and Y

ii) The conditional probability distribution of X given Y = 1

3. If X and Y are independent with

$p_X(0) = 0.5,\ p_X(1) = 0.3,\ p_X(2) = 0.2$ and $p_Y(0) = 0.6,$

$p_Y(1) = 0.1(2) = p_Y(3) = 0.05$ and $p_Y(4) = 0.2.$

Display the joint pmf of (X, Y).

Answers:

1. (i) $P_X(100) = 0.50, P_X(250) = 0.50, P_Y(0) = 0.25, P_Y(100) = 0.25, P_Y(200) = 0.50$

(ii)$P(Y \geq 100) = 0.75$, (iii)X and Y are not independent

(iv)$P(Y = 100 \mid X = 100) = 0.20$, $P(Y = 100 \mid X = 250) = 0.30$

2. (i) $P_X(0) = \frac{1}{3}, P_X(1) = \frac{2}{3}, P_Y(-1) = \frac{1}{3}, P_Y(1) = \frac{2}{3}$

(ii) $P(X = 0 \mid Y = 1) = \frac{1}{2}, P(X = 1 \mid Y = 1) = \frac{1}{2}$

3) Joint pmf of (X, Y)

$X\backslash Y$	0	1	2	3	4
0	0.30	0.025	0.025	0.025	0.10
1	0.18	0.015	0.015	0.015	0.06
2	0.12	0.010	0.010	0.010	0.04

Joint Distribution (Continuous rv 's)

- Let X and Y be continuous rv's. A joint probability density function $f(x, y)$ for these two variables is a function satisfying

$$f(x, y) \geq 0$$

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

Also, $P(a \le X \le b,\ c \le Y \le d) = \int_a^b \int_c^d f(x,\ y)\, dy\, dx$

- The marginal probability density function of X and Y, denoted by $f_X(x)$ and $f_Y(y)$ are given by

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

$f_Y(y) = \int_{-\infty}^{\infty} f(x,\ y)\, d$x for $-\infty <$ Y $< \infty$

- Two continuous random variables *X* and *Y* are said to be **independent** if for every pair of *x* and *y* values

$$f(x,\ y) = f_X(x) \cdot f_Y(y)$$

- If it is not satisfied, then X and Y are said to be dependent.
- Conditional Distributions Let X and Y be two continuous rv's with joint pdf $f(x,\ y)$ and marginal probability distribution of X is $f_X(x)$.
- Then for any X value of x for which $f_X(x) > 0$, the conditional probability density function of Y when X = x is $f_{Y|X}(y|x) = \frac{f(x,\ y)}{f_X(x)}$

Joint Probability Distribution Function

Let (X, Y) be a two-dimensional continuous random variable then their joint distribution function is given by

$$F_{XY}(x,\ y) = P(X \le x,\ Y \le y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(x,\ y) dx dy$$

Marginal Distribution Function

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

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

1) If the joint probability density of two random variables is given by

$$f(x_1, x_2) = 6e^{-2x_1-3x_2}, x_1 > 0,\ x_2 > 0$$

$$= 0, \qquad \text{elsewhere}$$

Find the probabilities that

a) The first random variable will take on value between 1and 2 and the second random variable will take on a value between 2 and 3.

b) The first random variable will take on a value less than 2 and the second random variable will take on a value greater than 2.

Sol. a) $\int_1^2 \int_2^3 6\, e^{-2x_1-3x_2}\, dx_1 dx_2 = (e^{-2} - e^{-4})(e^{-6} - e^{-9})$

$= 0.0003$

b) $\int_0^2 \int_2^\infty 6\, e^{-2x_1-3x_2}\, dx_1 dx_2 = (1 - e^{-4})(e^{-6})$

$= 0.0025$

2) If the joint probability density of two random variables is given by

f(x , y) = c(2x+y) for 2<x<6, 0<y<5

= 0 else where.

Find i) C ii) P(x>3,y>2) iii) P(x + y < 4)

Sol:

(i) $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) dydx = 1$

$$\int_2^6 \int_0^5 \mathrm{c(2x + y)dydx} = 1$$

$$\int_2^6 c\left(2xy + \frac{y^2}{2}\right)_0^5 dx = 1$$

$$\int_2^6 c\left(10x + \frac{25}{2}\right) dx = 1$$

$$c\left(\frac{10x^2}{2} + \frac{25x}{2}\right)_2^6 = 1$$

$$210c = 1$$

$$c = \frac{1}{210}$$

(ii) $P(x > 3, y > 2) = \frac{1}{210}\int_3^6 \int_2^5 (2\mathrm{x} + \mathrm{y})\mathrm{dydx} = \frac{15}{28}$

(iii) $P(x + y < 4) = \frac{1}{210}\int_2^4 \int_0^{4-x} (2\mathrm{x} + \mathrm{y})\mathrm{dydx} = \frac{2}{35}$

3) If the joint probability density of two random variables is given by

$f(x_1, x_2) = 6e^{-2x_1-3x_2}$ for $x_1>0$, $x_2>0$

= 0 , elsewhere

Find the joint distribution function of the two random variables. Also find the probability that both random variables will take on values less than 1.

Sol. Joint distribution function is,

$F(x_1,x_2) = \int_o^{x_1} \int_0^{x_2} 6\, e^{-2x_1-3x_2}\, dx_1 dx_2$ for $x_1>0$, $x_2>0$

= 0, else where

So that $F(x_1, x_2) = (1 - e^{-2x_1})(1 - e^{-3x_2})$ for $x_1>0$, $x_2>0$

= 0; else where

The probability that both random variables will take on values less than 1 is

$F(1,1) = \int_o^1 \int_0^1 6\, e^{-2x_1-3x_2}\, dx_1 dx_2$

$= (1 - e^{-2})(1 - e^{-3}) \quad = \quad 0.8216$

Exercise

1) Let the joint pdf for (X, Y) be

$$f(x, y) = \begin{cases} 24xy, & 0 \le x \le 1,\ 0 \le y \le 1, x + y \le 1 \\ 0, & otherwise \end{cases}$$

Find (i) Is f(x, y) legitimate pdf.

(ii) To compute the probability that the two types of nuts together make up at most 50% of the can.

2) If the joint probability density of two random variables is given by

f(x , y) = bxy for 1<x<3, 2<y<4

= 0, else where.

Find i. b

ii. Marginal density function

3) If the joint probability density of two random variables is given by

$$f_{XY}(x, y) = \begin{cases} x^2 + \dfrac{xy}{3} & 0 < x < 1, 0 < y < 2 \\ & 0, otherwise \end{cases}$$

Find conditional density function of x for given y.

Answers:

1. (i) Yes, (ii) $\frac{1}{16} = 0.0625$

2. (i) $b = \frac{1}{24}$, (ii) $f_X(x) = \begin{cases} \frac{x}{4}, & 1 < x < 3 \\ 0, & \text{otherwise} \end{cases}$ and $f_Y(y) = \begin{cases} \frac{y}{6}, & 2 < y < 4 \\ 0, & \text{otherwise} \end{cases}$

3. $f_{X|Y}(x \mid y) = \frac{6x^2 + 2xy}{2+y}, 0 < x < 1,\ 0 < y < 2$

UNIT 3

Discrete and Continuous distributions

In this unit, we will discuss some of the probability distributions. The values of the random variables and their corresponding probabilities can be arranged so that the probabilities are represented as a mathematical function of the random variable values.

Probability distributions are broadly classified based on the type of random variable they describe. The two main categories are:

1. Discrete Probability distributions

2. Continuous Probability distributions.

In this present unit, we shall discuss two important discrete probability distributions, Binomial and Poisson distributions. And in continuous probability distributions, we shall discuss about Uniform and Normal distributions.

The other discrete distributions are Bernoulli, Geometric, Negative binomial and Hyper-geometric distributions and other continuous distributions are Exponential, Gamma, Beta and Weibull distributions.

Binomial Distribution:

The Binomial Distribution is a discrete probability distribution that gives the probability of obtaining a fixed number of successes in a fixed number of independent trials.

Definition: A random variable X follows a binomial distribution if:

- There are n independent trials

- Each trial has only two outcomes: success or failure
- Probability of success is constant, say p

  This can be denoted by X~Binomial(n, p)

A random variable X has a binomial distribution if it assumes only non-negative values and its probability mass function is given by $P(X = x) = nC_x p^x q^{n-x}, for\ x = 0,1,2,\dots n$ where n is number of trials, x is number of success, p is probability of success.

Examples:

1. Number of Students passed in an examination.
2. Number of defective items in a batch.
3. Reliability of a machine.
4. Number of companies achieved sales targets.

Conditions of Binomial distribution:

A random variable follows a **Binomial Distribution** only when the following conditions are satisfied:

1. Trials are repeated under similar conditions for a fixed number of times i.e., n times.
2. Each trial has only two possible outcomes, either success or failure.
3. The probability of success in each trial should be constant.
4. The trials are independent.

Measures of Binomial distribution

Mean of the Binomial distribution:

The binomial probability distribution is given by $p(x) = P(X = x) = nC_x p^x q^{n-x}, for\ x = 0, 1, 2, \dots n$

Mean of X is given by $\mu = E(X) = \sum_{x=0}^{n} x\, p(x)$

$$= 0.q^n + 1.nC_1 pq^{n-1} + 2.nC_2 p^2 q^{n-2} + \dots\dots + n.p^n$$

$$= npq^{n-1} + 2.\frac{n(n-1)}{2!} p^2 q^{n-2} + 3.\frac{n(n-1)(n-2)}{3!} p^3 q^{n-3} + \dots + np^n$$

$$= np\left[q^{n-1} + (n-1)pq^{n-2} + \frac{(n-1)(n-2)}{2!}p^2q^{n-3} + \ldots + p^{n-1}\right]$$
$$= np(q+p)^{n-1}$$
$$= np(1) = np$$

Variance of the Binomial distribution:

Variance of X is given by $V(X) = E(X^2) - (E(X))^2 = \sum_{x=0}^{n} x^2\, p(x) - \mu^2$

$$= \sum_{x=0}^{n} (x(x-1) + x)\, p(x) - \mu^2$$

$$= \sum_{x=0}^{n} x(x-1)p(x) + \sum_{x=0}^{n} xp(x) - \mu^2$$

$$= [2.nC_2p^2q^{n-2} + 3.2.nC_3p^3q^{n-3} + \ldots .n(n-1)p^n] + \mu - \mu^2$$

$$= \left[2.\frac{n(n-1)}{2!}p^2q^{n-2} + 6.\frac{n(n-1)(n-2)}{3!}p^3q^{n-3} + \cdots + n(n-1)p^n\right] + \mu - \mu^2$$

$$= n(n-1)p^2[q^{n-2} + (n-2)pq^{n-3} + \ldots + p^{n-2}] + \mu - \mu^2$$

$$\therefore V(X) = n(n-1)p^2(q+p)^{n-2} + \mu - \mu^2$$
$$= n(n-1)p^2(1)^{n-2} + \mu - \mu^2$$

$$= n^2p^2 - np^2 + np - n^2p^2 = np(1-p) = npq \; since\; \mu = np$$

Mode of the Binomial distribution:

Mode of the binomial distribution is the value of x at which $p(x)$ has maximum value.

$$\text{Mode} = \begin{cases} \text{integral part of } (\mathrm{n}+1)\mathrm{p}, \text{if } (\mathrm{n}+1)\mathrm{p} \text{ is not an integer} \\ (\mathrm{n}+1)\mathrm{p} \text{ and } (\mathrm{n}+1)\mathrm{p}-1, \text{if } (\mathrm{n}+1)\mathrm{p} \text{ is an integer} \end{cases}$$

Recurrence Relation for the Binomial distribution:

Recurrence relation is a relation in which probability of $r+1$ is expressed in terms probability of r. The recurrence relation for the binomial distribution is given by

$$P(r+1) = \frac{(n-r)p}{(r+1)q} P(r)$$

Solved Problems

1. A factory produces items with a 10% defect rate. If 4 items are selected, find the probability that exactly 1 is defective.

Sol. Let p be probability of defective item, $\text{p} = 10\% = 0.1$

and q be the probability of failure, then $\text{q} = 0.9$

here $\text{n} = 4,$ Let x denotes the number of successes (in this case is a defective item)

the probability that exactly 1 is defective in 4 items is given by $\text{P(X} = 1)$

By definition, $P(X = x) = nC_x p^x q^{n-x}, for\ x = 0,1,2,\ldots n$

Substituting $\text{n} = 4,\ \text{p} = 0.1, \text{q} = 0.9$ and $\text{x} = 1$ since we require exactly 1 is defective.

$$P(X = 1) = 4C_1(0.1)^1(0.9)^{4-1} = 0.2916$$

2. A student guesses answers to 6 multiple-choice questions (each has 4 options). Find the probability of getting exactly 2 correct.

Sol. Let p be probability of guessing correct answer, $\text{p} = \frac{1}{4} = 0.25$

and q be the probability of failure, then $\text{q} = 0.75$

here $\text{n} = 6,$ Let x denotes the number of successes (in this case is getting correct) the probability of getting exactly 2 correct in 6 questions is given by $P(X = 2)$

By definition, $P(X = x) = nC_x p^x q^{n-x}, for\ x = 0,1,2, \dots n$

Substituting $n = 6$, $p = 0.25, q = 0.75$ and $x = 2$ since we require exactly 2 correct.

$$P(X = 2) = 6C_2(0.25)^1(0.75)^{6-2} = 0.2966$$

3. A machine produces bolts with 95% accuracy. In a batch of 8 bolts, find the probability that at- least 7 are good.

Sol. Let p be probability of producing accurate bolt, p = 95% = 0.95

and q be the probability of getting defective bolt, then $q = 0.05$

Since batch of 8 bolts are considered, here $n = 8$,

Let x denotes the number of successes (in this case is a non-defective bolt)

The probability of at least 7 are good in 8 bolts is given by $P(X \geq 7) = P(X = 7) + P(X = 8)$

By definition, $P(X = x) = nC_x p^x q^{n-x}, for\ x = 0,1,2, \dots n$

Substituting $n = 8$, $p = 0.95, q = 0.05$ and $x = 7, x = 8$ since we require $P(X \geq 7)$.

$$P(X \geq 7) = P(X = 7) + P(X = 8)$$

$$= 8C_7(0.95)^7(0.05)^{8-1} + 8C_8(0.95)^8(0.05)^{8-8}$$

$$= 0.2793 + 0.6634 = 0.9427$$

4. A system has 6 independent components, each working with probability 0.8. The system works if at least 5 components work. Find the system reliability.

Sol. Let p be probability that the independent component work, $p = 0.8$

and q be the probability that the independent component fail to work, then $q = 0.2$

Since system has 6 independent components, consider $n = 6$,

Let x denotes the number of successes (in this case is independent component work)

The probability of at least 5 components will work in 6 independent components is given by $P(X \geq 5) = P(X = 5) + P(X = 6)$

By definition, $P(X = x) = nC_x p^x q^{n-x}, for\ x = 0,1,2, \ldots n$

Substituting $n = 6$, $p = 0.8, q = 0.2$ and $x = 5, x = 6$ since we require $P(X \geq 5)$.

$$P(X \geq 5) = P(X = 5) + P(X = 6)$$

$$= 6C_5(0.8)^5(0.2)^{6-5} + 6C_6(0.8)^6(0.2)^{6-6}$$

$$= 0.393216 + 0.262144 = 0.65536$$

5. A player has a 75% success rate in free throws. If he attempts 10 shots, find the probability that he makes (a) exactly 7 shots (b) at least 8 shots.

Sol. Number of shots: n = 10

Probability of success (making a shot): p = 0.75

Probability of failure: q = 0. 25

Let x denotes the number of successes (in this case is successful shot)

By definition, $P(X = x) = nC_x p^x q^{n-x}, for\ x = 0,1,2, \ldots n$

Probability of exactly 7 shots = $P(X = 7) = 10C_7(0.75)^7(0.25)^{10-7} = 0.2503$

Probability of at least 8 shots = $P(X \geq 8) = P(X = 8) + P(X = 9) + P(X = 10)$

$$= 10C_8(0.75)^8(0.25)^{10-8} + 10C_9(0.75)^9(0.25)^{10-9} + 10C_{10}(0.75)^{10}(0.25)^{10-10}$$

$$= 0.2816 + 0.1877 + 0.0563 = 0.5256$$

There is about 52.56% chance he makes 8 or more shots.

6. Four coins are tossed 160 times. The number of times x heads occur is given below.

x	0	1	2	3	4
Number of times	8	34	69	43	6

Fit a binomial distribution to this data on the assumption that coins are unbiased.

Sol. The coin is unbiased, probability of getting head is $p = \frac{1}{2}$ and $q = \frac{1}{2}$

$$N = \sum f_i = 8 + 34 + 69 + 43 + 6 = 160$$

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

and the recurrence relation is given by

$$P(r+1) = \frac{(n-r)p}{(r+1)q} P(r)$$

here $n = 4\ and\ r = 0,1,2,3,4$

$\frac{p}{q} = \frac{1/2}{1/2} = 1,$

$$P(0) = P(X = 0) = 4C_0 \left(\frac{1}{2}\right)^0 \left(\frac{1}{2}\right)^4 = \left(\frac{1}{2}\right)^4 for\ r = 0$$

No. of heads R	Observed frequency f	Probability p(x)	Expected or Theoretical frequency f(x) = Np(x)
0	8	$P(0) = P(X = 0) = \left(\frac{1}{2}\right)^4$	$f(0) = 160.P(0) = 160\left(\frac{1}{2}\right)^4 = 10$

1	34	$P(1) = \left(\frac{4-0}{0+1}\right) P(X = 0)$ $= \frac{4}{1}\left(\frac{1}{2}\right)^4 = \frac{1}{4}$	$f(1) = 160.P(1) = 160\left(\frac{1}{4}\right)$ $= 40$
2	69	$P(2) = \left(\frac{4-1}{1+1}\right) P(X = 1)$ $= \frac{3}{2}.\frac{1}{4} = \frac{3}{8}$	$f(2) = 160.P(2) = 160\left(\frac{3}{8}\right)$ $= 60$
3	43	$P(3) = \left(\frac{4-2}{2+1}\right) P(X = 2)$ $= \frac{2}{3}.\frac{3}{8} = \frac{1}{4}$	$f(3) = 160.P(3) = 160\left(\frac{1}{4}\right)$ $= 40$
4	6	$P(4) = \left(\frac{4-3}{3+1}\right) P(X = 3)$ $= \frac{1}{4}.\frac{1}{4} = \frac{1}{16}$	$f(4) = 160.P(4) = 160\left(\frac{1}{16}\right)$ $= 10$

7. Five dice are thrown 243 times. How many times do you expect at least two dice to show a 4 or 5?

Sol. p = Probability of occurrence of 4 or 5 in one throw $= \frac{2}{6} = \frac{1}{3}$

$$\therefore q = 1 - p = 1 - \frac{1}{3} = \frac{2}{3}, n = 5$$

Let x denotes the number of successes (in this case is two dice to show a 4 or 5)

The probability of getting at least two dice to show a 4 or 5

$$= P(X \geq 2) = P(X = 2) + P(X = 3) + P(X = 4) + P(X = 5)$$
$$= 5_{C_2}\left(\frac{1}{3}\right)^2\left(\frac{2}{3}\right)^3 + 5_{C_3}\left(\frac{1}{3}\right)^3\left(\frac{2}{3}\right)^2 + 5_{C_4}\left(\frac{1}{3}\right)^4\left(\frac{2}{3}\right)^1 + 5_{C_5}\left(\frac{1}{3}\right)^5$$
$$= \frac{1}{3^5}[10 \cdot 4 \cdot 2 + 10 \cdot 2 \cdot 4 + 5 \cdot 2 + 1]$$
$$= \frac{1}{243}(80 + 80 + 10 + 1) = \frac{171}{243}$$

The expected number of such cases in 243 times

$$= 243\left(\frac{171}{243}\right) = 171$$

Chebyshev's Theorem:

Let X be a random variable with mean μ and standard deviation. Then for any positive number k, the probability that value of X lies in the interval $(\mu - k\sigma, \mu + k\sigma)$ is $\geq 1 - \frac{1}{k^2}$

That is, $P[\mu - k\sigma \leq X \leq \mu + k\sigma] \geq 1 - \frac{1}{k^2}$. Or $P(|X - \mu| \leq k\sigma) \geq 1 - \frac{1}{k^2}$

Its complement is, probability that X deviates from mean μ by atleast $k\sigma$ is atmost $\frac{1}{k^2}$.

That is, $P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$

Solved Problems

1. A random variable X has a mean $\mu = 10$ and a variance $\sigma^2 = 9$. Using Chebyshev's theorem, find

a) $P(4 < X < 16)$ b) $P(|X - 10| \geq 6)$

Sol. Given $\mu = 10, \sigma = 3;$

By Chebyshev's theorem $P(\mu - k\sigma < X < \mu + k\sigma) \geq 1 - \frac{1}{k^2}$

a) $As\ we\ have\ \mu + k\sigma = 16\ and\ \mu - k\sigma = 4,$

$on\ simplification\ we\ get\ k = 2$

$$P(4 < X < 16) \geq 1 - \frac{1}{2^2} = \frac{3}{4}$$

b) we have by complement of Chebyshev's theorem,

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$$

$$P(|X - 10| \geq 6) \leq \frac{1}{2^2} = \frac{1}{4}$$

2. Suppose X is a random variable with mean $\mu = 100$ and standard deviation $\sigma = 10$.

a) Find the conclusion one can derive from Chebyshev's inequality for $k = 2$ and $k = 3$.

b) Find an interval $[a, b]$ about the mean $\mu = 100$ for which the probability that X lies in the interval is at least 99 percent.

Sol. a) Setting $k = 2$, we get

$\mu - k\sigma = 100 - 2(10) = 80, \qquad \mu + k\sigma = 100 + 2(10) = 120,$

$1 - \frac{1}{k^2} = 1 - \frac{1}{2^2} = \frac{3}{4}.$

Thus, from Chebshev's inequality we can conclude that probability that X lies between 80 and 120 is at least $\frac{3}{4}$. that is $P(80 < X < 120) \geq 1 - \frac{1}{2^2} = \frac{3}{4}$

Similarly for $k = 3$

$$\mu - k\sigma = 100 - 3(10) = 70, \qquad \mu + k\sigma = 100 + 3(10) = 130,$$

we get $P(70 \leq X \leq 130) \geq 1 - \frac{1}{3^2} = \frac{8}{9}.$

b) Here we set, $1 - \frac{1}{k^2} = 0.99$ and Solve for k. This yields

$1 - 0.99 \geq \frac{1}{k^2}$ or $0.01 = \frac{1}{k^2}$ this implies $k = 10$.

Thus the desired interval is,

$$[a, b] = [\mu - k\sigma, \mu + k\sigma] = [100 - 10(10), 100 + 10(10)] = [0,200].$$

3. Find the probability, using Chebyshev's theorem that the number of driving licences X issued by Road Transport Authority (R T A) in a specific month is between 64 and 184 if the numbers of driving licenses issued is a random variable with $\mu = 124$ and $\sigma = 7.5$.

Sol. $X = \mu - k\sigma$

for $X = 64$, $64 = 124 - k\,(7.5)$

$k = 8$

$$X = \mu + k\sigma$$

for $X = 184$, $184 = 124 + k\,(7.5)$

$k = 8$

By Chebyshev's theorem $P(\mu - k\sigma < X < \mu + k\sigma) \geq 1 - \frac{1}{k^2}$

or $P(|X - \mu| \leq k\sigma) \geq 1 - \frac{1}{k^2}$

$$P(64 < X < 184) \geq 1 - \frac{1}{8^2} = 0.984$$

Exercise

1. Suppose a random variable X has the mean $\mu = 25$ and standard deviation $\sigma = 2$. Use Chebshev's inequality to estimate (a) $P(X \leq 35)$ (b) $P(X \geq 20)$.

2. Let X be random variable with mean $\mu = 40$ and $\sigma = 5$. Find the value of for which $(40-b \leq X \leq 40+b) \geq 0.95$.

Answers

1. a) 0.96   b) 0.84

2. 23.4

Poisson Distribution:

Poisson distribution – Definition

A random variable follows X a Poisson distribution if:

- The experiment counts the number of occurrences of an event in a fixed interval (time, area, volume, etc.)

- The events occur independently
- The average rate of occurrence is constant, say λ
- Two events cannot occur simultaneously in a very small interval

This can be denoted by X~Poisson (λ)

A random variable X has a Poisson distribution if it assumes only non-negative integer values 0,1,2,.. and its probability mass function is given by

$$P(X = x) = \frac{e^{-\lambda}\lambda^{x}}{x!} \; for \; x = 0,1,2\ldots$$

Where λ = average number of occurrences (mean rate), x = number of occurrences and e = Euler's constant (≈ 2.718)

Application of Poisson Distribution

Poisson distribution could be used to predict:

- Text messages per hour
- Machine malfunctions per year
- Website visitors per month
- Influenza cases per year
- Number of Telephone calls per minute
- Number of misprints per page etc.

Mean of the Poisson distribution:

Let X ~ Poisson(λ)

$$P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}, x = 0,1,2,\ldots$$

Mean of Poisson distribution $\mu = E(X) = \sum_{x=0}^{\infty} x \cdot P(X = x)$

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

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

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

Let $y = x - 1$

$$\mu = E(X) = e^{-\lambda} \sum_{y=0}^{\infty} \frac{\lambda^{y+1}}{y!} = \lambda e^{-\lambda} \sum_{y=0}^{\infty} \frac{\lambda^y}{y!} = \lambda e^{-\lambda} e^{\lambda} = \lambda$$

$$\therefore \mu = \lambda$$

Variance of the Poisson distribution:

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

First find $E[X(X-1)]$

$$E[X(X-1)] = \sum x(x-1)P(X = x)$$

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

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

Let $y = x - 2$

$$E[X(X-1)] = e^{-\lambda}\sum_{y=0}^{\infty}\frac{\lambda^{y+2}}{y!} = \lambda^2 e^{-\lambda}\sum_{y=0}^{\infty}\frac{\lambda^{y}}{y!} = \lambda^2$$

Now,

$$E(X^2) = E[X(X-1)] + E(X)$$
$$= \lambda^2 + \lambda$$
$$\sigma^2 = E(X^2) - [E(X)]^2$$
$$= (\lambda^2 + \lambda) - \lambda^2$$

$$\sigma^2 = \lambda$$

Solved Problems

1. 3% of the items of a factory are defective. The items are packed in a box of 200 items. What is the probability that there will be
   (i) 1 defective item
   (ii) at most two defective items.

Sol.

Given p = probability of defective item = 3% = 0.03 and $n = 200$

$$\therefore \text{Mean} = \lambda = np = (0.03)(200) = 6$$

Let x denotes the number of successes (in this case it is defective item)

Now the Poisson distribution is

$$P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!} = \frac{e^{-6}6^x}{x!}$$

(i) $p(x = 1) = \frac{e^{-6}6^1}{1!} = 6e^{-6} = 0.0149$

(ii) $p(\text{at most } 2) = p(x \le 2)$

$$= p(0) + p(1) + p(2)$$
$$= e^{-6}\left[\frac{6^0}{0!} + \frac{6^1}{1!} + \frac{6^2}{2!}\right]$$
$$= e^{-6}[1 + 6 + 18] = 25e^{-6} = 0.0619$$

2. The average number of accidents on a highway is 2 per day. Find the probability that on a given day there are
(i) exactly 2 accidents
(ii) more than 2 accidents.

Sol. Average number of accidents on a highway $\lambda = 2$

Let x denotes the number of successes (in this case it is number of accidents on a highway)

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

(i) $p(x = 2) = \frac{e^{-2}2^{2}}{2!} = \frac{4}{2}e^{-2} = 2e^{-2} = 0.2707$

(ii) p(more than 2) $= p(x > 2) = 1 - [p(0) + p(1) + p(2)]$

$$= 1 - e^{-2}\left[\frac{2^0}{0!} + \frac{2^1}{1!} + \frac{2^2}{2!}\right]$$
$$= 1 - e^{-2}(1 + 2 + 2) = 1 - 5e^{-2}$$
$$= 1 - 0.6767 = 0.3233$$

3. A typist makes on an average 3 errors per page. Find the probability that a page contains
(i) no errors
(ii) at most one error.

Sol. The average number of errors per page $\lambda = 3$

Let x denotes the number of successes (in this case it is errors per page)

$$P(X = x) = \frac{e^{-3}3^{x}}{x!}$$

(i) $p(X = 0) = e^{-3} = 0.0498$

(ii) p(at most 1) = p(x ≤ 1)

$$= e^{-3}\left[\frac{3^0}{0!}+\frac{3^1}{1!}\right]$$

$$= e^{-3}(1+3) = 4e^{-3} = 0.1992$$

4. The average number of calls received at a call center is 5 per hour. Find the probability that in a given hour there are
(i) exactly 4 calls
(ii) at least 3 calls.

Sol. The average number of calls received at a call center is $\lambda = 5$ per hour

Let x denotes the number of successes (in this case it is number of calls received at a call center)

$$P(X = x) = \frac{e^{-5}5^x}{x!}$$

(i) $p(x = 4) = \frac{e^{-5}5^4}{4!} = \frac{625}{24}e^{-5} = 0.1755$

(ii) p(at least 3) = p(x ≥ 3) = $1 - e^{-5}\left[\frac{5^0}{0!}+\frac{5^1}{1!}+\frac{5^2}{2!}\right]$

$$= 1 - e^{-5}(1+5+12.5) = 1 - 18.5e^{-5}$$

$$= 1 - 0.1247 = 0.8753$$

5. A book contains misprints with an average of 1.5 per page. Find the probability that a page contains
(i) exactly 1 misprint
(ii) not more than 2 misprints.

Sol. Average number of misprints in a book is $\lambda = 1.5$ per page.

Let x denotes the number of successes (in this case it is number of misprints in a book)

$$P(X = x) = \frac{e^{-1.5}(1.5)^x}{x!}$$

(i) $p(x = 1) = \frac{e^{-1.5}(1.5)^1}{1!} = 1.5e^{-1.5} = 0.3347$

(ii) p(not more than 2) = $p(x \leq 2)$

$$= e^{-1.5}\left[\frac{(1.5)^0}{0!} + \frac{(1.5)^1}{1!} + \frac{(1.5)^2}{2!}\right]$$
$$= e^{-1.5}(1 + 1.5 + 1.125) = 3.625e^{-1.5}$$
$$= 0.8088$$

6. If X follows Poisson distribution and $P(X = 2) = P(X = 3)$, find the mean and variance of the distribution.

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

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

Cancel $e^{-\lambda}$:

$$\frac{\lambda^2}{2} = \frac{\lambda^3}{6}$$
$$3\lambda^2 = \lambda^3$$
$$\lambda = 3$$
$$\therefore \text{Mean} = \lambda = 3$$
$$\therefore \text{Variance} = \lambda = 3$$

7. If the ratio $\frac{P(X = 1)}{P(X = 3)} = \frac{3}{2}$. Find the mean and variance.

Sol. Given $\frac{P(X = 1)}{P(X = 3)} = \frac{3}{2}$

$$\frac{\frac{e^{-\lambda}\lambda^1}{1!}}{\frac{e^{-\lambda}\lambda^3}{3!}} = \frac{3}{2}$$
$$\frac{\lambda}{\lambda^3/6} = \frac{3}{2}$$
$$\frac{6}{\lambda^2} = \frac{3}{2}$$

$$12 = 3\lambda^2$$
$$\lambda^2 = 4 \Rightarrow \lambda = 2$$

∴ Mean $= 2$ and Variance = 2.

8. If $P(X = 0) + P(X = 1) = 2P(X = 2)$. Find the mean and variance.

Sol. Given $P(X = 0) + P(X = 1) = 2P(X = 2)$

$$P(0) = e^{-\lambda},$$

$$P(1) = \lambda e^{-\lambda},$$

$$P(2) = \frac{\lambda^2}{2} e^{-\lambda}$$
$$e^{-\lambda} + \lambda e^{-\lambda} = 2 \cdot \frac{\lambda^2}{2} e^{-\lambda}$$

Cancel $e^{-\lambda}$:
$$1 + \lambda = \lambda^2$$
$$\lambda^2 - \lambda - 1 = 0$$
$$\lambda = \frac{1 + \sqrt{5}}{2} \approx 1.618$$

∴ Mean $= 1.618$

∴ Variance $= 1.618$

9. Fit a Poisson distribution to the following data and find the expected frequencies.

x (No. of defects)	0	1	2	3	4
f (Frequency)	30	50	40	20	10

Sol.

$$N = \sum f = 30 + 50 + 40 + 20 + 10 = 150$$
$$\sum fx = (0)(30) + (1)(50) + (2)(40) + (3)(20) + (4)(10)$$
$$= 0 + 50 + 80 + 60 + 40 = 230$$

$$\therefore \lambda = \frac{\sum fx}{\sum f} = \frac{230}{150} = 1.53$$

Now

$$P(X = x) = \frac{e^{-1.53}(1.53)^x}{x!}$$

Using recurrence:

$$P(0) = e^{-1.53} = 0.216$$
$$P(1) = P(0)\frac{1.53}{1} = 0.216 \times 1.53 = 0.330$$
$$P(2) = P(1)\frac{1.53}{2} = 0.252$$
$$P(3) = P(2)\frac{1.53}{3} = 0.129$$
$$P(4) = P(3)\frac{1.53}{4} = 0.049$$

Expected frequencies: $E(x) = NP(x)$

X	P(x)	Expected Frequency
0	0.216	32.4
1	0.33	49.5
2	0.252	37.8
3	0.129	19.35
4	0.049	7.35

Exercise

1. The average number of defects in a fabric roll is 2 per roll. Find the probability that a roll has
   (i) no defects
   (ii) at most 2 defects.

2. A telephone exchange receives on an average 4 calls per minute. Find the probability that in a given minute there are
   (i) exactly 3 calls
   (ii) more than 2 calls.

3. The average number of misprints on a page is 1. Find the probability that a page contains
   (i) exactly 2 misprints
   (ii) at least one misprint.

4. The number of accidents in a factory follows Poisson distribution with mean 3 per week. Find the probability that in a week there are
   (i) no accidents
   (ii) at most 1 accident.

5. A typist makes on an average 2 errors per page. Find the probability that a page contains
   (i) exactly 1 error
   (ii) not more than 2 errors.

6. The average number of customers arriving at a shop is 5 per hour. Find the probability that in a given hour there are
   (i) exactly 6 customers
   (ii) at least 4 customers.

Answers:

1 (i) 0.1353 (ii) 0.6767

2 (i) 0.1954 (ii) 0.7619

3(i) 0.1839 (ii) 0.6321

4(i) 0.0498 (ii) 0.1992

5(i) 0.2707 (ii) 0.6767

6(i) 0.1462 (ii) 0.7350

Poisson Process

A Poisson process is a model for a series of discrete events where the average time between events is known, but the exact timing of events is random. Its criteria is

- Events are independent of each other.
- The average rate (events per time period) is constant.
- Two events cannot occur at the same time

If m is the average number of outcomes per unit time then the probability of number of outcomes occurring in a given time interval t, is given by

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

Where $\lambda = m * t$

Solved Problems

1. A person receives on average 3 e-mails per hour.
   a) What is the probability that he will receive 5 e-mails over a period two hours?
   b) What is the probability that he will receive more than 2 e-mails over a period two hours?

Sol.

We are given the average per hour $m = 3$. We therefore need to find the average λ over a period of two hours.

$\lambda = 3 * 2 = 6$, e-mails over 2 hours

a) The probability that he will receive 5 e-mails over a period two hours is given by the Poisson probability formula

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

$$P(x = 5) = \frac{e^{-6}(6)^5}{5!} = 0.1606.$$

b) More than 2 e-mails,

$$P(X > 2) = P(x = 3) + p(x = 4) + p(x = 5) + \cdots$$

$$P(X > 2) = 1 - \big(P(X = 0) + P(X = 1) + p(X = 2)\big)$$

$$P(X > 2) = 1 - \left(\frac{e^{-6}6^0}{0!} + \frac{e^{-6}6}{1!} + \frac{e^{-6}6^2}{2!}\right) = 0.938$$

Exercise

1. If a bank receives on the average 6 bad cheques per day, what are the probabilities that it will receive
   (a) 4 bad cheques on any given day?
   (b) 10 bad cheques over any 2 consecutive days?
2. During inspection of continuous process of making large rolls of floor coverings, 0.5 imperfections are spotted per minute on average. Use the Poisson distribution to find the probabilities

(a) One imperfection in 4 minutes.

(b) At least two in 8 minutes

(c) At most one in 10 minutes

Answers

1. a) 0.134 b) 0.105

2. a) 0.2707 b) 0.9085 c) 0.0404

Normal Distribution

A Normal Distribution (also called Gaussian distribution) is a continuous probability distribution that is symmetric about its mean, where most of the observations cluster around the central peak and probabilities for values further away from the mean taper off equally in both directions.

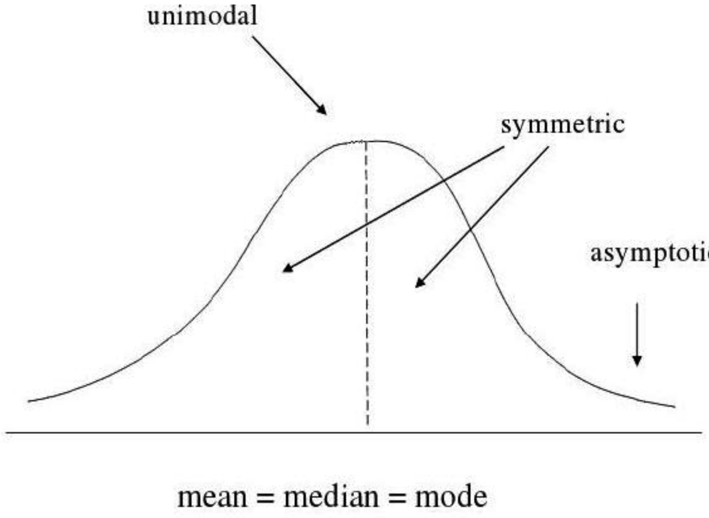
Mathematical Definition

A random variable X is said to follow a normal distribution if its probability density function (PDF) is:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, -\infty < x < \infty$$

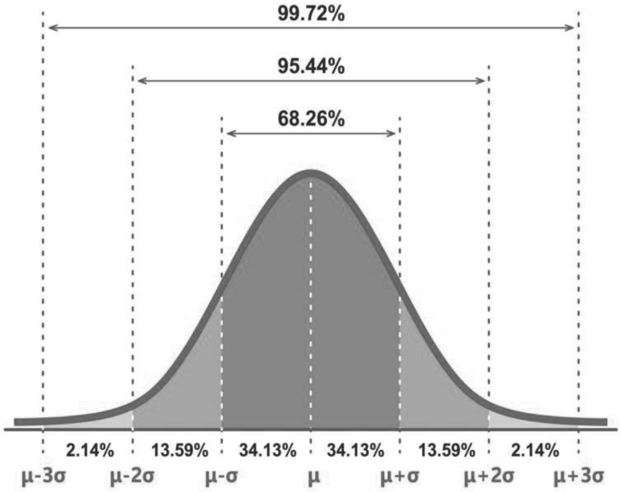
where:

- μ = Mean
- σ = Standard deviation
- σ^2 = Variance



Key Properties of Normal distribution:

1. Shape and Symmetry: The normal distribution has a bell-shaped curve. It is perfectly symmetric about its center. The left and right halves are mirror images.
2. Mean, Median and Mode: In a normal distribution, Mean = Median = Mode. All three lie at the center (peak) of the curve. This indicates no skewness.
3. Total Area Under the Curve: The total area under the curve is 1 (or 100%). This represents the total probability. The curve extends from $-\infty$ to $+\infty$, but most values cluster near the mean.
4. Empirical Rule (68–95–99.7 Rule): About 68% of data lies within 1 standard deviation (σ) of the mean. About 95% lies within 2σ. About 99.7% lies within 3σ.



5. Defined by Two Parameters: A normal distribution is completely determined by:
   - Mean (μ) → determines the location (center)
   - Standard deviation (σ) → determines the spread (width)
6. Standard Normal Distribution: A special case where Mean $\mu = 0$ and Standard deviation $\sigma = 1$. Values are converted using $Z = \frac{X-\mu}{\sigma}$. Used for probability calculations via Z-tables.
7. Skewness and Kurtosis: Skewness = 0 → perfectly symmetric and Kurtosis = 3 → called mesokurtic (moderate peak).
8. Infinite Tails (Asymptotic Nature): The curve extends indefinitely but never touches the x-axis. Extreme values are possible but very rare.

Examples of Normal distribution:

1. Heights of People

- Heights of adults in a population often follow a normal distribution
- Most people are of average height, few are very short or very tall

2. Marks in an Examination

- In many exams, student marks form a bell-shaped curve
- Most students score **average marks**, few score very high or very low

3. Measurement Errors

- Errors in scientific measurements tend to be normally distributed
- Small errors occur frequently, large errors are rare

Measures of Normal distribution

Consider a random variable $X \sim N(\mu, \sigma^2)$with probability density function:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, -\infty < x < \infty$$

1. Mean of Normal Distribution:

The mean is defined as:

$$E(X) = \int_{-\infty}^{\infty} xf(x)\, dx$$

Substitute f(x):

$$E(X) = \int_{-\infty}^{\infty} x \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$

Transformation

Let:

$$z = \frac{x-\mu}{\sigma} \Rightarrow x = \mu + \sigma z, dx = \sigma dz$$

Then:

$$E(X) = \int_{-\infty}^{\infty} (\mu + \sigma z)\frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz$$

Split the integral:

$$E(X) = \mu \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz + \sigma \int_{-\infty}^{\infty} z \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz$$

Since $\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz = 1$

and $\int_{-\infty}^{\infty} z \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz = 0$(odd function)

we get $E(X) = \mu(1) + \sigma(0) = \mu$

2. Variance of Normal Distribution:

Variance is:

$$\begin{aligned} Var(X) &= E[(X - \mu)^2] \\ &= \int_{-\infty}^{\infty} (x - \mu)^2 f(x)\, dx \end{aligned}$$

Substitute PDF:

$$= \int_{-\infty}^{\infty} (x - \mu)^2 \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$

Using substitution:

$z = \frac{x-\mu}{\sigma} \Rightarrow (x - \mu) = \sigma z, dx = \sigma dz$

$$\begin{aligned} Var(X) &= \int_{-\infty}^{\infty} (\sigma z)^2 \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz \\ &= \sigma^2 \int_{-\infty}^{\infty} z^2 \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz \end{aligned}$$

Standard Result:

$$\int_{-\infty}^{\infty} z^2 \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz = 1$$

Thus:

$$Var(X) = \sigma^2$$

Mode of a Normal Distribution:

Consider a continuous random variable X following a normal distribution with mean μ and standard deviation σ. Its probability density function is:

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

Step 1: Definition of Mode

The mode is the value of x at which the function $f(x)$ attains its maximum value.

Step 2: First Derivative

To find the maximum, differentiate $f(x)$ with respect to x:

$$\frac{df(x)}{dx} = \frac{1}{\sigma\sqrt{2\pi}} \cdot \frac{d}{dx} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Using chain rule:

$$\frac{df(x)}{dx} = \frac{1}{\sigma\sqrt{2\pi}} \cdot \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \cdot \left(-\frac{2(x-\mu)}{2\sigma^2}\right)$$

$$\frac{df(x)}{dx} = \frac{1}{\sigma\sqrt{2\pi}} \cdot \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \cdot \left(-\frac{x-\mu}{\sigma^2}\right)$$

Step 3: Critical Point

Set the derivative equal to zero:

$$\frac{df(x)}{dx} = 0$$

Since exponential terms are never zero, we get:

$$x - \mu = 0$$

$$\Rightarrow x = \mu$$

Step 4: Second Derivative Test

To confirm maximum, compute second derivative:

$$\frac{d^2 f(x)}{dx^2} < 0 \text{at } x = \mu$$

This shows that $f(x)$ has a maximum at $x = \mu$.

Median of a Normal Distribution

Consider a continuous random variable $X \sim N(\mu, \sigma^2)$with probability density function:

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

Step 1: Definition of Median

The median M is the value that satisfies:

$$P(X \leq M) = \frac{1}{2}$$

That is, $\int_{-\infty}^{M} f(x)\,dx = \frac{1}{2}$

Step 2: Standardization

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

Then, $P(X \leq M) = P\left(Z \leq \frac{M-\mu}{\sigma}\right)$

So,

$$\int_{-\infty}^{M} f(x)\,dx = \int_{-\infty}^{\frac{M-\mu}{\sigma}} \frac{1}{\sqrt{2\pi}} e^{-z^2/2}\,dz$$

Step 3: Use Symmetry of Standard Normal Distribution

The standard normal distribution is symmetric about $z = 0$, hence:

$$P(Z \leq 0) = \frac{1}{2}$$

So we must have $\frac{M-\mu}{\sigma} = 0$

Step 4: Solve for Median

$$M - \mu = 0$$
$$\Rightarrow M = \mu$$

For a normal distribution:

$$\text{Mean} = \text{Median} = \text{Mode} = \mu$$

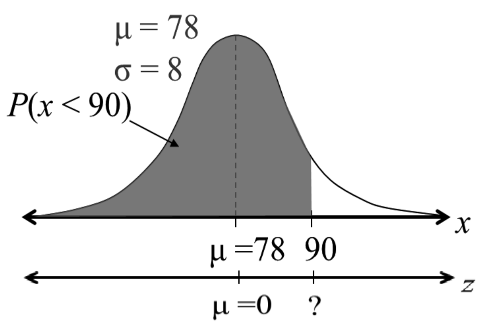
This occurs because the normal curve is perfectly symmetric about its mean.

Solved Problems

1. The average on a Machine Learning test was 78 with a standard deviation of 8. If the test scores are normally distributed, find the probability that a student receives a test score less than 90.

Sol. Let $X \sim N(\mu, \sigma^2)$

Given average on a Machine Learning was 78 with a standard deviation of 8



To find $P(x < 90)$

$$Z = \frac{90 - \mu}{\sigma} = \frac{90 - 78}{8} = 1.5$$

$P(x < 90) = P(z < 1.5) = 0.5 + P(z = 1.5)$

$= 0.5 + 0.4332$

$= 0.9332$

The probability that a student receives a test score less than 90 is 0.9332.

2. For an exam, average score was 78 with a standard deviation of 8. If the exam scores are normally distributed, find the probability that a student receives a test score between 60 and 80.

Sol. Let $X \sim N(\mu, \sigma^2)$

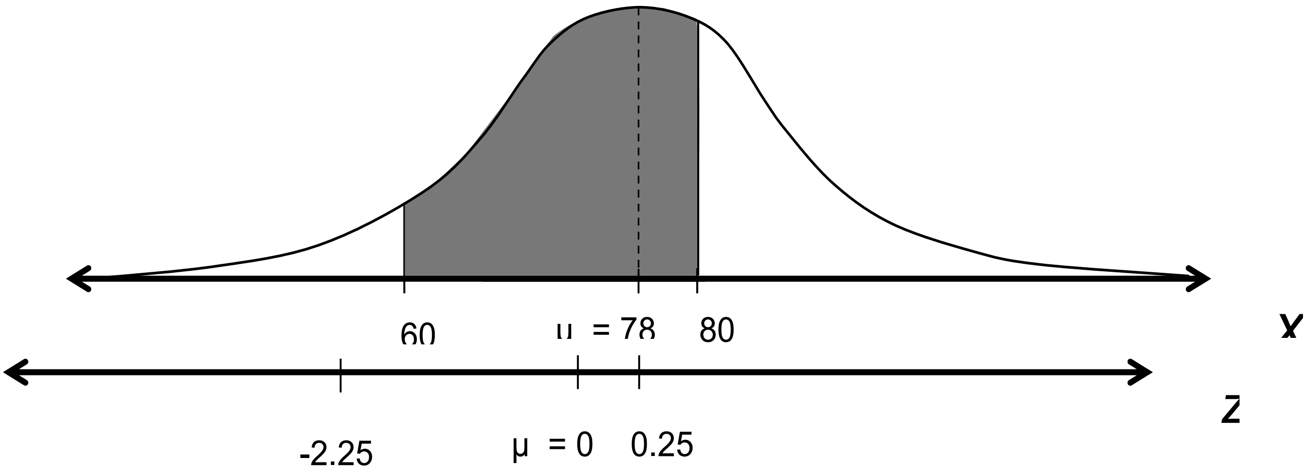
Given $\mu = 78, \sigma = 8$

To find, $P(60 < X < 80)$

Convert to Z:

$$Z_1 = \frac{60-\mu}{\sigma} = \frac{60-78}{8} = -2.25$$

$$Z_2 = \frac{80-\mu}{\sigma} = \frac{80-78}{8} = 0.25$$



$$P(60 < X < 80 = P(-2.25 < Z < 0.25)$$

$$= 0.5865$$

The probability that a student receives a test score between 60 and 80 is 0.5865.

3. In a normal distribution, given mean is 16.2 and variance is 1.5625. Find the probabilities that X takes on value a) >16.8 b) <14.9 c) between 13.6 and 18.8.

Sol. Let $X \sim N(\mu, \sigma^2)$

Given $\mu = 16.2, \sigma = 1.25$

$a) \, P(X > 16.8)$

$$z = \frac{x-\mu}{\sigma} = \frac{16.8-16.2}{1.25} = 0.48$$

$$P(X > 16.8) = P(Z > 0.48) = 1 - P(Z < 0.48) = 0.3156$$

$b) \, P(X < 14.9)$

$$z = \frac{x-\mu}{\sigma} = \frac{14.9-16.2}{1.25} = -1.04$$

$$P(X < 14.9) = P(Z < -1.04) = 0.149$$

$c) \, P(13.6 < X < 18.8)$

$$z_1 = \frac{x-\mu}{\sigma} = \frac{13.6-16.2}{1.25} = -2.08$$

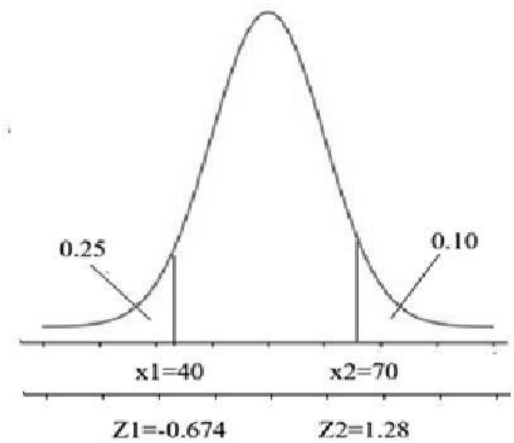
$$z_2 = \frac{x-\mu}{\sigma} = \frac{18.8-16.2}{1.25} = 2.08$$

$$P(13.6 < X < 18.8) = P(Z < 2.08) - P(Z < -2.08)$$
$$= 0.9812 - 0.0188 = 0.9624$$

4. In a normal distribution, 25% of the items are less than 40 and 10% are greater than 70. Find the mean and standard deviation.

Sol. Let $X \sim N(\mu, \sigma^2)$
Given:

$$P(X < 40) = 0.25, P(X > 70) = 0.10$$



Convert to Z:

$$Z_1 = \frac{40-\mu}{\sigma}, P(Z < Z_1) = 0.25 \Rightarrow Z_1 = -0.674$$
$$Z_2 = \frac{70-\mu}{\sigma}, P(Z < Z_2) = 0.90 \Rightarrow Z_2 = 1.28$$

$$\frac{40-\mu}{\sigma} = -0.674 \qquad (1)$$
$$\frac{70-\mu}{\sigma} = 1.28 \qquad (2)$$

Solve:

$$(70-\mu) - (40-\mu) = 1.28\sigma + 0.674\sigma$$
$$30 = 1.954\sigma \Rightarrow \sigma = 15.35$$

Substitute:

$$40 - \mu = -0.674(15.35) \approx -10.35$$
$$\mu = 50.35$$

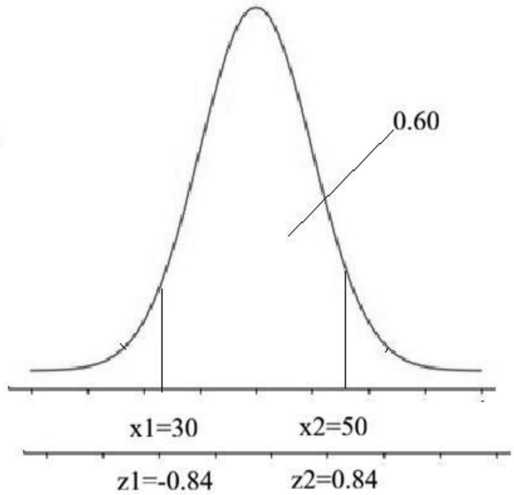
5. In a normal distribution, 60% of the items lie between 30 and 50, and the distribution is symmetric. Find mean and standard deviation.

Sol.
Since symmetric:

$$\mu = \frac{30 + 50}{2} = 40$$

Given:

$$P(30 < X < 50) = 0.60$$



So:

$$P(-z < Z < z) = 0.60 \Rightarrow P(0 < Z < z) = 0.30$$

From tables:

$$z = 0.84$$

Thus:

$$\frac{50 - 40}{\sigma} = 0.84 \Rightarrow \frac{10}{\sigma} = 0.84$$
$$\sigma = 11.90$$

6. In a normal distribution, 20% of items are less than 35 and 50% of items are less than 50. Find mean and standard deviation.

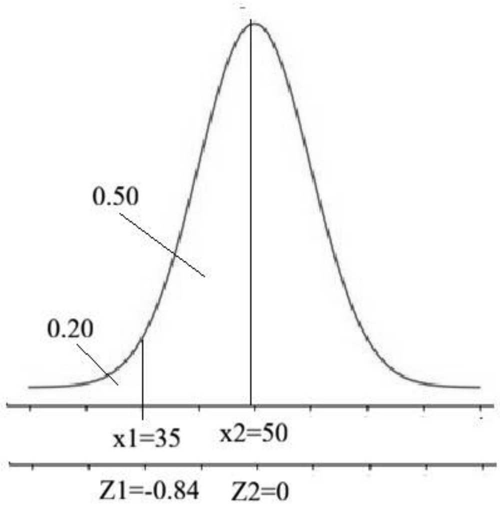
Sol.

Since:

$$P(X < 50) = 0.5 \Rightarrow \mu = 50$$

Given:

$$P(X < 35) = 0.20$$
$$Z = \frac{35 - 50}{\sigma} = \frac{-15}{\sigma}$$



From table:

$$P(Z < -0.84) = 0.20 \Rightarrow Z = -0.84$$

Thus:

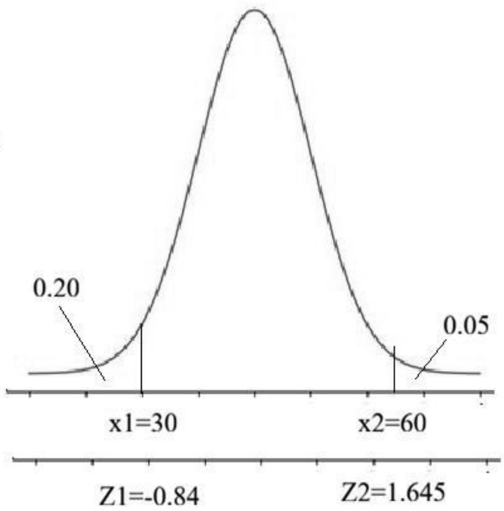
$$\frac{-15}{\sigma} = -0.84 \Rightarrow \sigma = \frac{15}{0.84} = 17.86$$

7. In a normal distribution, 20% of the items are less than 30 and 5% are greater than 60. Find the mean and standard deviation.

Sol. Given:

$$P(X < 30) = 0.20, P(X > 60) = 0.05$$

Let $X \sim N(\mu, \sigma^2)$



Step 1: Convert to Z

$$Z_1 = \frac{30 - \mu}{\sigma}, P(Z < Z_1) = 0.20 \Rightarrow Z_1 = -0.84$$
$$Z_2 = \frac{60 - \mu}{\sigma}, P(Z < Z_2) = 0.95 \Rightarrow Z_2 = 1.645$$

Step 2: Form equations

$$\frac{30 - \mu}{\sigma} = -0.84(1)$$
$$\frac{60 - \mu}{\sigma} = 1.645(2)$$

Step 3: Solve

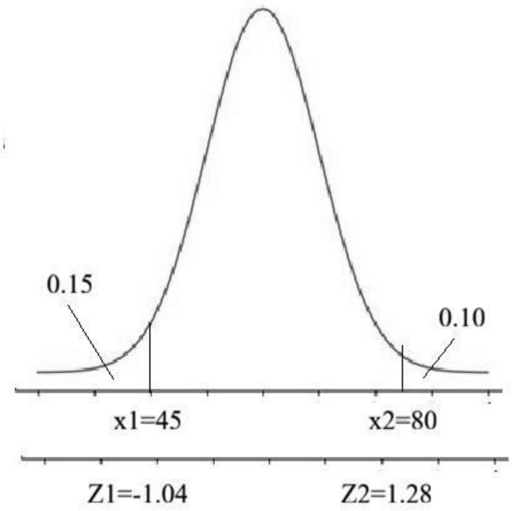
$$30 = (1.645 + 0.84)\sigma = 2.485\sigma \Rightarrow \sigma = 12.07$$
$$30 - \mu = -0.84(12.07) = -10.14 \Rightarrow \mu = 40.14$$

Answer:

$$\boxed{\mu \approx 40.14, \sigma \approx 12.07}$$

7. In a normal distribution, 15% of the items are less than 45 and 10% are greater than 80. Find the mean and standard deviation.
Sol.
Given

$$P(X < 45) = 0.15$$
$$P(X > 80) = 0.10$$



$$Z_1 = -1.04, Z_2 = 1.28$$

$$\frac{45-\mu}{\sigma} = -1.04 \qquad (1)$$

$$\frac{80-\mu}{\sigma} = 1.28 \qquad (2)$$

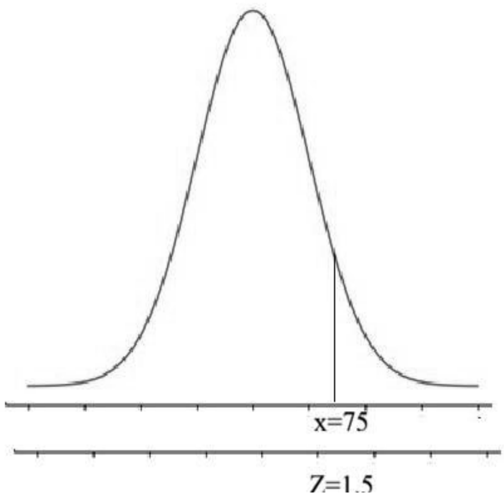
$$35 = (1.28 + 1.04)\sigma = 2.32\sigma \Rightarrow \sigma = 15.09$$

$$45 - \mu = -1.04(15.09) = -15.69 \Rightarrow \mu = 60.69$$

8. Marks of 800 students are normally distributed with mean $\mu = 60$ and standard deviation $\sigma = 10$. Find:
(i) Number of students scoring above 75
(ii) Highest marks of the lowest 20% students
(iii) Limits of the middle 60% students

Sol.

(i) $P(X > 75)$

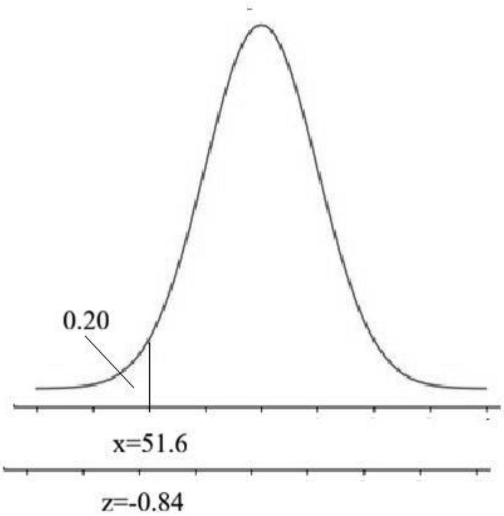


$$Z = \frac{75 - 60}{10} = 1.5$$

$P(Z < 1.5) = 0.9332 \Rightarrow P(X > 75) = 1 - 0.9332 = 0.0068$

Students $= 0.0668 \times 800 = 53$

(ii) Lowest 20%

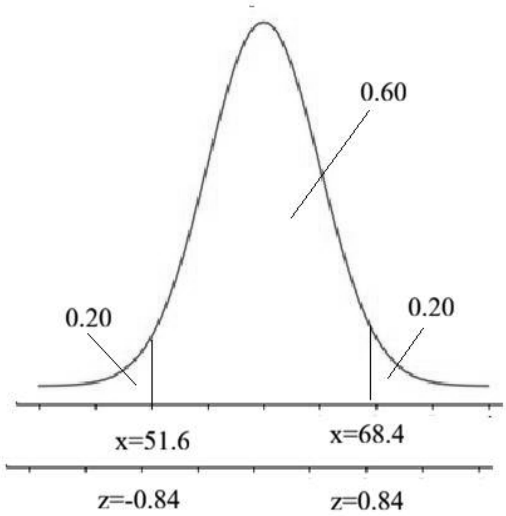


$P(Z < z) = 0.20 \Rightarrow z = -0.84$

$x = \mu + z\sigma = 60 - 0.84(10) = 51.6$

(iii) Middle 60%

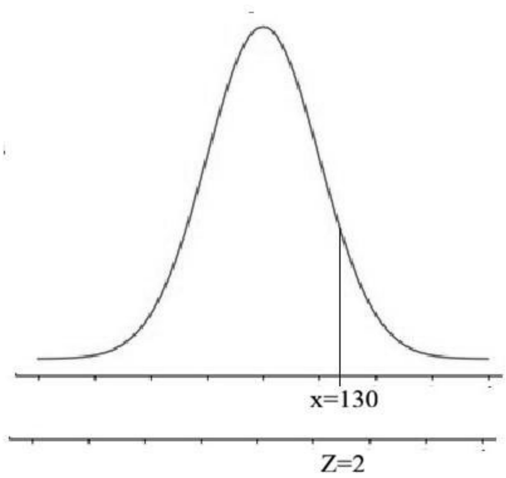
Remaining tails = 40% → each tail = 20%



$$z = \pm 0.84$$
$$x = 60 \pm 0.84(10) = 60 \pm 8.4$$
Limits: (51.6, 68.4)

9. IQ scores of 1000 students follow $N(100, 15^2)$. Find:
(i) Number scoring above 130
(ii) Lowest 5% cutoff
(iii) Range of middle 90%

Sol (i) Number scoring above 130

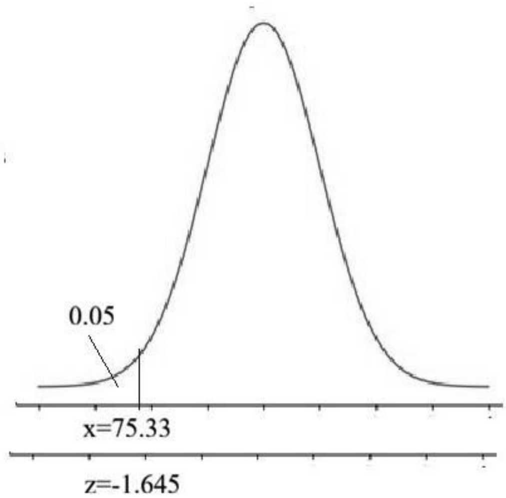


$$Z = \frac{130 - 100}{15} = 2$$

$$P(X > 130) = 1 - 0.9772 = 0.0228$$

$$= 0.0228 \times 1000 = 23$$

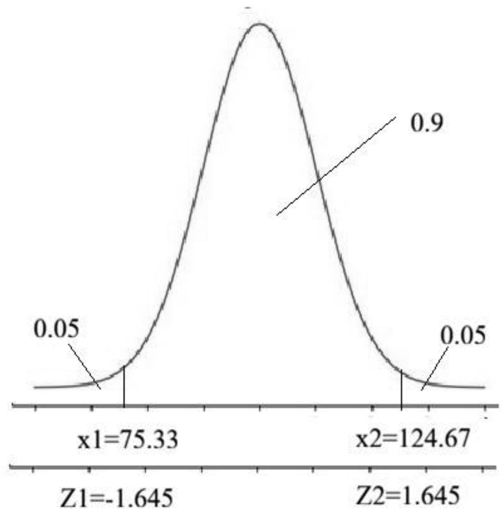
(ii) Lowest 5% cutoff



$$P(Z < z) = 0.05 \Rightarrow z = -1.645$$

$$x = 100 - 1.645(15) = 75.33$$

(iii) Middle 90% → tails 5% each

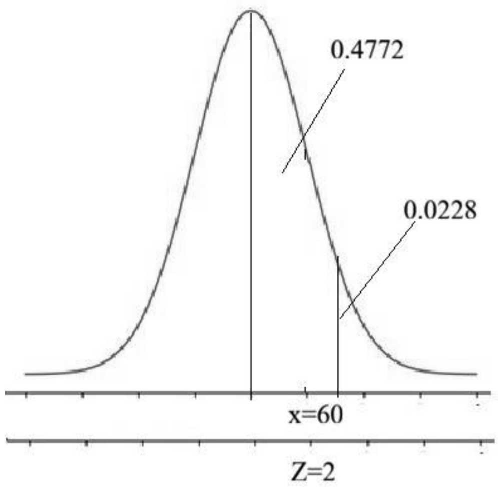


$$z = \pm 1.645$$

$$x = 100 \pm 1.645(15) = 100 \pm 24.675$$

The x lies between $(75.33\,,\ 124.67)$

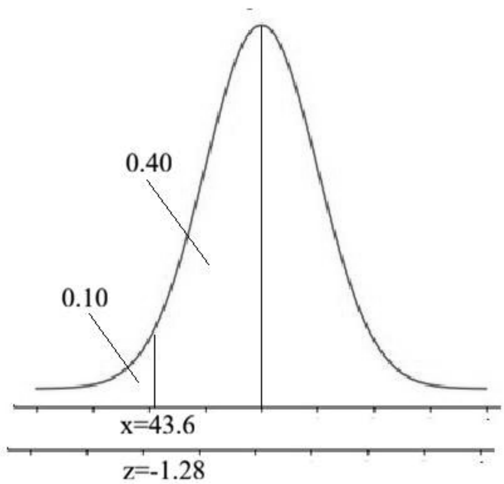
10. Scores of 600 students follow $N(50, 5^2)$. Find the following for the given information:
(i) Number scoring above 60
(ii) Lowest 10% cutoff
(iii) Middle 80% limits
Sol. **(i)** Number scoring above 60



For $X = 60$, we get $Z = \frac{60-50}{5} = 2$

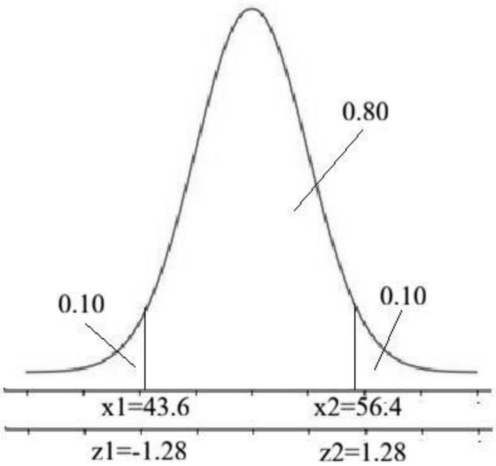
$$P(X > 60) = P(Z > 2) = 0.5 - P(0 < Z < 2) = 0.5 - 0.4772 = 0.0228$$

The number of students scoring above 60 is $= P(Z > 2)x600 = 0.0228 \times 600 = 14$
(ii) Lowest 10% cutoff



$$z = -1.28$$
$$x = 50 - 1.28(5) = 43.6$$

iii) Middle 80% limits



$P(Z < z1) = 0.10$ then by using Z-tables, we get z1 = -1.28
$P(Z > z2) = 0.10$ then by using Z-tables, we get z2 = 1.28

For $z = \pm 1.28$, $x = 50 \pm 6.4$
The range of x is $(43.6\,, 56.4)$.

Exercise:

1. In a normal distribution, 30% of the items are less than 45 and 15% are greater than 75. Find the mean and standard deviation.
2. In a normal distribution, 25% of the items are less than 32 and 5% are greater than 68. Find the mean and standard deviation.
3. In a normal distribution, 10% of the items are greater than 90 and 20% are less than 50. Find the mean and standard deviation.
4. In a normal distribution, 40% of the items lie between 48 and 72 symmetrically. Find the mean and standard deviation.
5. In a normal distribution, 50% of the items are less than 65 and 20% are less than 45. Find the mean and standard deviation.
6. In a normal distribution, 15% of the items are less than 28 and 10% are greater than 64. Find the mean and standard deviation.
7. The marks of 1000 students are normally distributed with mean 60 and standard deviation 12. Find: (i) Number of students scoring above 72
(ii) Lowest 10% cutoff marks (iii) Limits of middle 80% students
8. IQ scores of 800 students are normally distributed with mean 100 and standard deviation 20. Find: (i) Number scoring above 130 (ii) Lowest 5% cutoff (iii) Range of middle 90%
9. The weights of 600 students are normally distributed with mean 55 kg and standard deviation 8 kg. Find: (i) Number weighing above 63 kg
(ii) Lowest 20% cutoff weight (iii) Middle 60% limits

10. The heights of 1200 students are normally distributed with mean 165 cm and standard deviation 10 cm. Find: (i) Number taller than 180 cm
(ii) Shortest height among top 10% students (iii) Middle 50% limits
11. A factory produces bulbs whose life is normally distributed with mean 900 hours and standard deviation 120 hours. Find: (i) Number lasting above 1050 hours out of 2000 bulbs (ii) Lowest 5% lifetime (iii) Middle 90% lifetime range
12. The scores of 500 students are normally distributed with mean 70 and standard deviation 15. Find: (i) Number scoring below 50 (ii) Highest marks of lowest 25% students (iii) Middle 70% limits

Answers:

1. μ = 55.08, σ = 19.23.	2. μ = 42.46.9, σ = 15.52	3. μ = 65.86, σ = 18.84
4. μ = 60, σ = 22.90	5. μ = 65, σ = 23.76	6. μ = 44.09, σ = 15.53
7. (i) 159 students,	(ii) 44.62,	(iii) 44.62 to 75.38
8. (i) 54 students,	(ii) 67.1,	(iii) 67.1 to 132.9
9. (i) 95 students,	(ii) 48.27 kg,	(iii) 48.27 kg to 61.73 kg
10. (i) 80 students,	(ii) 177.8 cm,	(iii) 158.25 cm to 171.75 cm
11. (i) 211 bulbs,	(ii) 702.6 hours	(iii) 702.6 hours to 1097.4 hours
12. (i) 46 students,	(ii) 59.88,	(iii) 54.45 to 85.55

Normal Approximation to a Binomial Distribution

If $np \geq 15$ and $nq \geq 15$, then the binomial random variable x is approximately normally distributed with mean $\mu = np, \sigma = \sqrt{npq}$
Then the standard normal random variable is $z = \frac{x-\mu}{\sigma} = \frac{x-np}{\sqrt{npq}}$

Correction for Continuity:

Since the binomial distribution is discrete and the normal distribution is continuous, a continuity correction is applied. This involves adjusting the discrete binomial probability by 0.5 in either direction to account for the continuous nature of the normal distribution.
When approximating

$(X \leq k)$, use $(X \leq k+0.5)$

$(X<k)$, use $(X<k-0.5)$

$(X \geq k)$, use $(X \geq k-0.5)$

$(X>k)$, use $(X>k+0.5)$.

Solved Problems

1. If 62% of all clouds seeded with silver iodide show spectacular growth, What is the probability that among 40 clouds seeded with silver iodide at most 20 will show spectacular growth?

Sol. $Here\ n = 40,\ \ p = 0.62,\ q = 0.38,$

$$\mu = np = 24.8,\ \ \sigma^2 = npq = 9.424,\ \ \sigma = 3.0699$$

$$np = 24.8 \geq 15\ \ and\, nq = 15.2 \geq 15$$

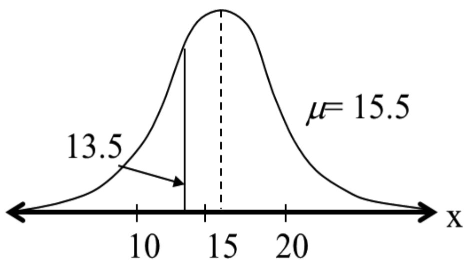
Therefore, we will use normal approximation to binomial distribution.

$$P(x \leq 20.5) = ?$$

$$When\ x = 20.5,\ Z = \frac{x-\mu}{\sigma} = \frac{20.5 - 24.8}{3.0699} = -1.40$$

$$\therefore\ P(x \leq 20.5) = P(Z \leq -1.40) = 0.0808$$

2. Thirty-one percent of the seniors in a certain high school plan to attend college. If 50 students are randomly selected, find the probability that less than 14 students plan to attend college.



Sol. $Here\ n = 50,\ \ p = 0.31,\ q = 0.69,$

$\mu = np = (50)(0.31) = 15.5,$

$\sigma^2 = npq = 10.6929,\ \ \sigma = 3.27$

$np = 15.5 \geq 15\ \ and\, nq = 34.5 \geq 15$

Therefore, we will use normal approximation to binomial distribution.

$$P(x \leq 13.5) = ?$$

$$When\ x = 13.5,\ Z = \frac{x-\mu}{\sigma} = \frac{13.5 - 15.5}{3.27} = -0.61$$

$$\therefore\ P(x \le 13.5) = P(Z \le -0.61)$$

$$= 0.2709$$

The probability that less than 14 plan to attend college is 0.2709.

Exercise

1. Find the probability that the proportion of heads will fall anywhere from 0.49 to 0.51 when a balanced coin is flipped 1000 times?

Answers

1. 0.4908

Uniform Distribution:

The continuous uniform distribution is one of the simplest probability distributions.
Definition
A random variable X is said to have a continuous uniform distribution over an interval [a, b] if its probability density function (PDF) is constant throughout that interval.
Mathematical Form

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

Explanation

- Every value between aand bis equally likely.
- The graph of the PDF is a horizontal line between a and b.
- The total area under the curve is 1 (as required for any probability distribution).

Key Properties

- Mean: $\frac{a+b}{2}$
- Variance: $\frac{(b-a)^2}{12}$

Example

If $X \sim U(0,10)$, then any value between 0 and 10 has the same likelihood of occurring.

1. Let $X \sim U(2,10)$. Find $P(4 \leq X \leq 7)$.

Sol.

$$P = \frac{7-4}{10-2} = \frac{3}{8} = 0.375$$

2. If $X \sim U(0,5)$, find the mean.

Sol.
Using the formula:

$$\mu = \frac{a+b}{2}$$
$$\mu = \frac{0+5}{2} = 2.5$$

3. Find variance for $X \sim U(1,7)$

Sol. Using the formula, $\mathrm{Var}(X) = \frac{(b-a)^2}{12}$

$$= \frac{(7-1)^2}{12} = \frac{36}{12} = 3$$

1. Mean (Expected Value)
The mean is defined as:

$$E(X) = \int_a^b x f(x)\, dx$$

Using the PDF of uniform distribution:

$$f(x) = \frac{1}{b-a}, a \leq x \leq b$$

Step-by-step derivation:

$$E(X) = \int_a^b x \cdot \frac{1}{b-a}\, dx$$

$$= \frac{1}{b-a}\int_a^b x\,dx$$
$$= \frac{1}{b-a}\left[\frac{x^2}{2}\right]_a^b$$
$$= \frac{1}{b-a}\cdot\frac{b^2-a^2}{2}$$
$$= \frac{1}{b-a}\cdot\frac{(b-a)(a+b)}{2}$$
$$= \frac{a+b}{2}$$

2. Variance

Variance is defined as:

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

Step 1: Find $E(X^2)$

$$E(X^2) = \int_a^b x^2 f(x)\,dx$$
$$= \int_a^b x^2 \cdot \frac{1}{b-a}\,dx$$
$$= \frac{1}{b-a}\int_a^b x^2\,dx$$
$$= \frac{1}{b-a}\left[\frac{x^3}{3}\right]_a^b$$
$$= \frac{1}{b-a}\cdot\frac{b^3-a^3}{3}$$
$$= \frac{1}{b-a}\cdot\frac{(b-a)(a^2+ab+b^2)}{3}$$
$$= \frac{a^2+ab+b^2}{3}$$

Step 2: Substitute into variance formula

$$\mathrm{Var}(X) = \frac{a^2+ab+b^2}{3} - \left(\frac{a+b}{2}\right)^2$$
$$= \frac{a^2+ab+b^2}{3} - \frac{a^2+2ab+b^2}{4}$$

Taking LCM (12):

$$= \frac{4(a^2+ab+b^2) - 3(a^2+2ab+b^2)}{12}$$

$$= \frac{(4a^2 + 4ab + 4b^2 - 3a^2 - 6ab - 3b^2)}{12}$$

$$= \frac{a^2 - 2ab + b^2}{12} = \frac{(b-a)^2}{12}$$

Exercise:

1. A random variable is uniformly distributed between 5 and 25. Find the probability that the variable takes a value between 10 and 18.

2. The lifetime of a device is uniformly distributed between 2 years and 10 years. Find the mean lifetime.

3. A continuous random variable is uniformly distributed between 4 and 16. Find the variance.
4. The waiting time for a bus is uniformly distributed between 0 and 30 minutes. Find the probability that the waiting time is more than 20 minutes.

Answers:

1. 0.4   2. 6 years   3. 12   4. 0.3333

UNIT 4

Sampling Distribution and Estimations

Sampling Distribution

Population: In statistics, the term population refers to the complete set of all observations or items that are of interest in a particular study. The population may be finite or infinite depending on the situation. The total number of elements in the population is denoted by **N**.

Sample: A sample is a smaller group selected from the population for the purpose of analysis. Instead of studying the entire population, conclusions about it are drawn based on the sample. The number of observations in the sample is denoted by **n**.

Different methods of sampling

Sampling methods are broadly classified into two categories: probability sampling and non-probability sampling. Each method has its own advantages depending on the nature of the population and the objective of the study.

I. Probability Sampling Methods

1. Random sampling or Probability sampling

Random sampling is a method in which each element of the population has an equal chance of being selected. The selection is made purely by chance, ensuring that there is no bias in choosing the sample. The sample obtained through this method is called a random sample.

If the population size is N and the sample size is n, then:

- Number of samples with replacement = N^n
- Number of samples without replacement = ${}^{N}C_{n}$

2. Stratified sampling or Stratified Random sampling

Stratified sampling is used when the population is not uniform and consists of different groups. In this method, the population is divided into smaller homogeneous groups called strata based on certain characteristics. A random sample is then taken from each stratum, and all these sub-samples are combined to form the final sample. This method ensures proper representation of all sections of the population.

3. Systematic Sampling or Quasi – Random Sampling

In systematic sampling, the elements of the population are arranged in a particular order, and samples are selected at regular intervals. First, a starting point is chosen randomly from the initial units, and then every k^{th} item is selected to form the sample. This method is simple and easy to apply when the population is large.

II. Non – Probability Sampling Methods

1. Purposive Sampling or Judgment Sampling

In this method, the selection of sample elements is based on the judgment or experience of the investigator. The researcher chooses those units which are believed to represent the population effectively. This method is useful when expert opinion is required, but it may involve some bias. For example, if a sample of 20 students is to be selected from a class of 100 to analyze the extra-curricular activities of the students, the investigator would select 20 students who, in his judgment, would represent the class.

2. Sequential Sampling

Sequential sampling involves selecting samples in stages, one after another. After analyzing each sample, a decision is made whether to accept, reject, or continue sampling. If the result is not clear, additional samples are taken until a final conclusion is reached. This method is commonly used in quality control and decision-making processes.

Classification of Samples

Samples can be classified based on their size into large samples and small samples.

1. **Large Sample:** A sample is considered to be large when the number of observations is 30 or more ($n \geq 30$). In such cases, statistical analysis becomes simpler as the sample tends to follow normal distribution due to the central limit theorem.

2. **Small Sample:** A sample is said to be small when the number of observations is less than 30 ($n < 30$). For small samples, special statistical techniques are required, as the results may not follow normal distribution.

Parameter and Statistics

Parameter: A parameter is a numerical measure that describes a characteristic of the entire population. Since it is based on all observations in the population, its value is fixed but usually unknown. Examples include population mean (μ) and population variance (σ^2).

Statistic: A statistic is a numerical measure calculated from a sample. It is used to estimate the corresponding population parameter. Unlike parameters, statistics may vary from sample to sample. Examples include sample mean ($\bar{x}$) and sample variance (s^2).

Sample Mean and Sample Variance

In statistical analysis, when a sample is drawn from a population, it is important to summarize the data using suitable numerical measures. Two of the most commonly used measures are the sample mean and the sample variance. These statistics help in understanding the central tendency and variability of the sample data.

Sample Mean: If $X_1, X_2, \ldots, X_n$ represent a random sample of size *n*, the sample mean is defined as the average of all observations in the sample.

Sample mean is given by: $\bar{X} = \sum_{i=1}^{n} \frac{X_i}{n}$

It provides a measure of the central value of the sample and is used as an estimate of the population mean.

Sample Variance:

If $X_1, X_2 \ldots\ldots, X_n$ represents a random sample of size n, the sample variance measures how much the observations vary from the sample mean.

Sample variance is given by: $s^2 = \sum_{i=1}^{n} \frac{(X_i - \overline{X})^2}{n-1}$

It indicates the spread or dispersion of the sample data around the mean.

Central Limit Theorem

The Central Limit Theorem states that when a large sample is drawn from a population, the distribution of the sample mean tends to follow a normal distribution, regardless of the shape of the original population.

If $\overline{x}$is the mean of a sample of size n drawn from a population with mean μ and standard deviation σ, then the standardized sample mean $Z = \frac{\overline{x}-\mu}{\frac{\sigma}{\sqrt{n}}}$ follows approximately a standard normal distribution with mean 0 and standard deviation 1, for sufficiently large values of n, i.e. N(Z;0,1) as $n \to \infty$

Standard Error (S.E.) of a Statistic

Standard error is a measure of the variability of a sample statistic from sample to sample. It indicates how much a statistic is expected to vary from the corresponding population parameter.

1) S.E. $(\bar{x}) = \frac{\sigma}{\sqrt{n}}$
2) S.E. of sample proportion $p = \sqrt{\frac{PQ}{n}}$ where Q = 1-P
3) S.E. $(\overline{x_1} - \overline{x_2}) = \sqrt{\frac{\sigma_1}{n_1} + \frac{\sigma_2}{n_2}}$ where $\overline{x_1}$ and $\overline{x_2}$ are the means of two random samples of sizes n_1 and n_2 drawn from two populations with S.D. σ_1 and σ_2 respectively.
4) S.E. $(p_1 - p_2) = \sqrt{\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}}$

Finite Population: Consider a finite population of size **N** with mean **μ** and standard deviation **σ**. Suppose all possible samples of size **n** are drawn without replacement from this population.

(i) The mean of the sampling distribution of means (for N>n) is $\mu_{\overline{X}} = \mu$

(ii) The variance of the sampling distribution of means is $\sigma_{\overline{X}}{}^2 = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)$

Note : The factor $\left(\frac{N-n}{N-1}\right)$ is known as the finite population **correction factor.**

Solved Problems

1. What is the value of correction factor if n = 8 and N = 150

Sol. Given N = the size of the finite population = 150

n = the size of the sample = 8

Correction factor = $\frac{N-n}{N-1} = \frac{150-8}{150-1} = \frac{142}{149} = 0.959$

2. What is the value of correction factor if n = 20 and N = 500

Sol. Given N = the size of the finite population = 500

n = the size of the sample = 20

Correction factor = $\frac{N-n}{N-1} = \frac{500-20}{500-1} = \frac{480}{499} = 0.962$

3. A population consists of five numbers 4, 7, 9, 10, and 15. Consider all possible samples of size two which can be drawn with replacement from this population. Find

(a) The mean of the population

(b) The S.D. of the population

(c) The mean of the sampling distribution of means and

(d) The S.D. of the sampling distribution of means (i.e., the standard error of means)

Sol.

(a) Mean of the population is given by $\mu = \frac{4+7+9+10+15}{5} = \frac{45}{5} = 9$

(b) Variance of the population σ^2 is given by $\sigma^2 = \frac{\sum(x_i-\mu)^2}{n}$

$$= \frac{(4-9)^2 + (7-9)^2 + (9-9)^2 + (10-9)^2 + (15-9)^2}{5}$$

$$= \frac{25+4+0+1+36}{5} = \frac{66}{5} = 13.2 \therefore \sigma = \sqrt{13.2} = 3.63$$

(c) Sampling with replacement (infinite population):

The total no. of samples with replacement is

$N^n = 5^2 = 25$ samples of size 2

Here N = population size and n = sample size listing all possible samples of size 2 from population 4, 7, 9, 10, 15 with replacement we get 25 samples

All possible samples:

$$\begin{Bmatrix} (4,4) & (4,7) & (4,9) & (4,10) & (4,15) \\ (7,4) & (7,7) & (7,9) & (7,10) & (7,15) \\ (9,4) & (9,7) & (9,9) & (9,10) & (9,15) \\ (10,4) & (10,7) & (10,9) & (10,10) & (10,15) \\ (15,4) & (15,7) & (15,9) & (15,10) & (15,15) \end{Bmatrix}$$

Now compute the arithmetic mean for each of these 25 samples. The set of 25 means $\overline{x}$ of these 25 samples, gives rise to the distribution of means of the samples known as sampling distribution of means.

The samples means are

$$\begin{Bmatrix} 4 & 5.5 & 6.5 & 7 & 9.5 \\ 5.5 & 7 & 8 & 8.5 & 11 \\ 6.5 & 8 & 9 & 9.5 & 12 \\ 7 & 8.5 & 9.5 & 10 & 12.5 \\ 9.5 & 11 & 12 & 12.5 & 15 \end{Bmatrix}$$

And the mean of sampling distribution of means is the mean of these 25 means.

$$\mu_{\bar{x}} = \frac{\text{Sum of all sample means}}{25} = \frac{225}{25} = 9$$

(d) The variance of the sampling distribution of means is obtained by subtracting the mean 9 from each number and squaring the result, adding all 25 members thus obtained, and dividing by 25

$$\sigma_{\bar{x}}^2 = \frac{\sum(\bar{x}-9)^2}{25} = 6.6$$
$$\therefore \sigma_{\bar{x}} = \sqrt{6.6} = 2.57$$

Verification formula $\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n} = \frac{13.2}{2} = 6.6$

$$\therefore \sigma_{\bar{x}} = \sqrt{6.6} = 2.57$$

4. A population consists of five numbers 4, 7, 9, 10, and 15. Consider all possible samples of size two which can be drawn without replacement from this population. Find

(a) The mean of the population

(b) The S.D. of the population

(c) The mean of the sampling distribution of means and

(d) The S.D. of the sampling distribution of means (i.e., the standard error of means)

Sol.

(a) Mean of the population is given by $\mu = \frac{4+7+9+10+15}{5} = \frac{45}{5} = 9$

(b) Variance of the population σ^2 is given by $\sigma^2 = \frac{\sum(x_i-\mu)^2}{n}$

$$= \frac{(4-9)^2 + (7-9)^2 + (9-9)^2 + (10-9)^2 + (15-9)^2}{5} = \frac{25+4+0+1+36}{5}$$

$$= \frac{66}{5} = 13.2 \therefore \sigma = \sqrt{13.2} = 3.63$$

(c) Sampling without replacement (infinite population):

The total no. of samples with replacement is

${}^{N}C_n = {}^{5}C_2 = 10$ samples of size 2

Here N = population size and n = sample size listing all possible samples of size 2 from population 4, 7, 9, 10, 15 with replacement we get 25 samples

All possible samples:

$$\left\{ \begin{matrix} (4,7), (4,9), (4,10), (4,15), \\ (7,9), (7,10), (7,15), \\ (9,10), (9,15), \\ (10,15) \end{matrix} \right\}$$

Now compute the arithmetic mean for each of these 10 samples. The set of 10 means $\overline{x}$ of these 10 samples, gives rise to the distribution of means of the samples known as sampling distribution of means.

The samples means are

$$\left\{ \begin{matrix} 5.5, \quad 6.5, \quad 7, \quad 9.5, \\ 8, \quad 8.5, \quad 11 \\ 9.5, \quad 12 \\ 12.5 \end{matrix} \right\}$$

And the mean of sampling distribution of means is the mean of these 10 means.

$$\mu_{\bar{x}} = \frac{\text{Sum of all sample means}}{10} = \frac{90}{10} = 9$$

(d) The variance of the sampling distribution of means is obtained by subtracting the mean 9 from each number and squaring the result, adding all 10 members thus obtained, and dividing by 10

$$\sigma_{\bar{x}}^2 = \frac{\sum(\bar{x}-9)^2}{10} = 4.95$$

$$\therefore \sigma_{\bar{x}} = \sqrt{4.95} = 2.23$$

Verification formula (Finite Population Correction)

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$

$$= \frac{3.63}{\sqrt{2}} \times \sqrt{\frac{5-2}{5-1}}$$

$$= 2.57 \times \sqrt{\frac{3}{4}} = 2.57 \times 0.866 = 2.23$$

5. The mean weight of students in a college is 60 kg and S.D. is 12. What is the probability that the mean weight of 36 students is less than 63 kgs.

Sol.

μ = mean of the population
= mean weight of students of a college = 60 kg
σ = S.D. of population = 12 kgs
n = sample size = 36
$\bar{x}$ = Mean of samples = 63 kgs

Now

$$z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} = \frac{63-60}{12/\sqrt{36}} = \frac{3}{12/6} = \frac{3}{2} = 1.5$$

$$\therefore P(\bar{x} \le 63) = P(z < 1.5) = 0.5 + A(1.5)$$

$$= 0.5 + 0.4332 = 0.9332$$

Thus, the probability that the mean weight of 36 students less than 63 is **0.9332**

6. A random sample of size 64 is taken from an infinite population having the mean $\mu = 50$ and the variance $\sigma^2 = 144$. What is the probability that $\bar{x}$ will be between 48 and 52.

Sol.

μ = mean of the population = 50
σ^2 = variance of population = 144 i.e. $\sigma = 12$
n = sample size = 64
$\bar{x}_1 = 48$

Now

$$z_1 = \frac{\bar{x}_1 - \mu}{\sigma/\sqrt{n}} = \frac{48-50}{12/\sqrt{64}} = \frac{-2}{12/8} = \frac{-2}{1.5} = -1.33$$

And when

$$\bar{x}_2 = 52$$

Now

$$z_2 = \frac{\bar{x}_2 - \mu}{\sigma/\sqrt{n}} = \frac{52 - 50}{12/\sqrt{64}} = \frac{2}{1.5} = 1.33$$

$$\therefore P(48 \leq \bar{x} \leq 52) = P(-1.33 \leq z \leq 1.33)$$
$$= A(1.33) + A(1.33) = 0.4082 + 0.4082 = 0.8164$$

Thus, the probability that the sample mean lies between 48 and 52 is **0.8164**

7**.** A random sample of size 64 is taken from an infinite population having the mean 45 and the S.D. 8. What is the probability that $\bar{x}$will be between 46 and 47.5.

Sol.

$$\mu = \text{mean of the population} = 45$$
$$\sigma = \text{S.D. of the population} = 8$$
$$n = \text{sample size} = 64$$
$$\bar{x}_1 = 46$$

Now

$$z_1 = \frac{\bar{x}_1 - \mu}{\sigma/\sqrt{n}} = \frac{46 - 45}{8/\sqrt{64}} = \frac{1}{8/8} = 1$$

And when

$$\bar{x}_2 = 47.5$$

Now

$$z_2 = \frac{\bar{x}_2 - \mu}{\sigma/\sqrt{n}} = \frac{47.5 - 45}{8/\sqrt{64}} = \frac{2.5}{1} = 2.5$$

$$\therefore P(46 \leq \bar{x} \leq 47.5) = P(1 \leq z \leq 2.5)$$
$$= A(2.5) - A(1) = 0.4938 - 0.3413 = 0.1525$$

Thus, the probability that the sample mean lies between 46 and 47.5 is **0.1525**

Exercise

1. What is the value of correction factor if n = 5 and N = 200

2. What is the value of correction factor if n = 10 and N = 1000

3. A population consists of five numbers 2, 3, 6, 8 and 11. Consider all possible samples of size two which can be drawn with replacement from this population. Find

(a) The mean of the population

(b) The S.D. of the population

(c) The mean of the sampling distribution of means and

(d) The S.D. of the sampling distribution of means (i.e., the standard error of means)

4. A population consists of five numbers 2, 3, 6, 8 and 11. Consider all possible samples of size two which can be drawn without replacement from this population. Find

(a) The mean of the population

(b) The S.D. of the population

(c) The mean of the sampling distribution of means and

(d) The S.D. of the sampling distribution of means (i.e., the standard error of means)

5. A population consists of 6 numbers 5, 10, 14, 18, 13, and 24. Consider all possible samples of size two which can be drawn without replacement from this population. Find

(a) The mean of the population

(b) The S.D. of the population

(c) The mean of the sampling distribution of means and

(d) The S.D. of the sampling distribution of means (i.e., the standard error of means)

6. A random sample of size 64 is taken from a normal population with $\mu = 51.4$ and $\sigma = 68$. What is the probability that the mean of the sample will (a) exceed 52.9, (b) fall between 50.5 and 52.3, (c) 50.6

7. The mean height of students in a college is 155cms and S.D. is 15. What is the probability that the mean height of 36 students is less than 157 cms.

8. A random sample of size 100 is taken from an infinite population having the mean $\mu = 76$ and the variance $\sigma^2 = 256$. What is the probability that $\overline{x}$ will be between 75 and 78.

9. A random sample of size 64 is taken from an infinite population having the mean 45 and the S.D. 8. What is the probability that x will be between 46 and 47.5

Answers:

1) 0.9799, 2) 0.991

3) (a) $\mu = 6$, (b) $\sigma = \sqrt{10.8} = 3.29$, (c) $\mu_{\overline{x}} = 6$, (d) $\sigma_{\overline{x}} = \sqrt{(10.8 / 2)} = \sqrt{5.4} = 2.32$

4) (a) $\mu = 6$, (b) $\sigma = 3.29$, (c) $\mu_{\overline{x}} = 6$, (d) $\sigma_{x^-} = 2.02$

5) (a) $\mu = 14$, (b) $\sigma = 5.97$, (c) $\mu_{\overline{x}} = 14$, (d) $\sigma_{\overline{x}} = 3.77$

6) (a) 0.429, (b) 0.084, (c) 0.463, 7) 0.7881

8) 0.628 0.1525

ESTIMATION

In many practical situations, it is not feasible to collect information from an entire population due to limitations of time, cost, or accessibility. Instead, a subset of the population, known as a sample, is selected for analysis. The main objective of statistical inference is to use this sample information to draw conclusions about the population.

Population characteristics such as the mean, variance, or proportion are usually unknown and are referred to as parameters. Since these parameters cannot be observed directly, they must be approximated using sample data. The process of obtaining such approximations is known as estimation.

Estimation plays a fundamental role in statistics, as it provides a systematic way to make informed decisions and predictions about a population based on limited data.

Estimate: An estimate is a numerical value obtained from sample data that is used to approximate an unknown population parameter. It provides a practical way to infer information about the population when complete data is not available.

Estimator: An estimator is a statistical rule, formula, or method used to compute an estimate of a population parameter based on sample observations. For example, the sample mean $\overline{x}$ is commonly used as an estimator of the population mean μ.

Types of Estimation:

In statistical analysis, the unknown characteristics of a population are assessed using sample data. Depending on how the approximation is expressed, estimation can be broadly classified into following types.

(a) Point Estimation: A point estimate provides a single numerical value as an approximation of an unknown population parameter. This value is calculated from the observed sample and serves as the best immediate guess of the parameter.

(b) Interval Estimation: An interval estimate specifies a range of values within which the population parameter is expected to lie, with a certain level of confidence. Instead of giving a single value, it provides both lower and upper bounds, thereby indicating the precision of the estimate.

(c) Maximum Likelihood Estimation (MLE): The method of Maximum Likelihood Estimation is the best and most popular one among all methods to obtain an almost good or best estimator for a population parameter. It is a method of obtaining an estimator which most (maximum) likely estimates the true value of the parameter.

Interval Estimation of μ: The interval estimates for the population mean μ can be expressed as: $\bar{x} \pm z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$ (or) $(\overline{x} - E_{Max}, \overline{x} + E_{Max})$

where

$\bar{x}$ = sample mean

σ = population standard deviation

n = sample size

$z_{\alpha/2}$ = standard normal value corresponding to the desired confidence level

and $E_{Max} = z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$

Interval Estimation of p (proportion): The interval estimate for the population proportion $p$is given by

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \text{ (or) } (p - E_{Max}, p + E_{Max})$$

where $E_{Max} = z_{\frac{\alpha}{2}}\sqrt{\frac{PQ}{n}}$

Here

$\hat{p}$ = sample proportion

n = sample size

$z_{\alpha/2}$ = standard normal value corresponding to the confidence level

Confidence Level

In interval estimation, it is important to specify how reliable the estimated interval is. This reliability is expressed in terms of the confidence level.

The confidence level is the probability that the interval constructed from sample data will contain the true value of the population parameter. It reflects the degree of certainty associated with the estimation procedure.

Commonly used confidence levels are 90%, 95%, and 99%.

Interpretation of Confidence Levels

- A 90% confidence level implies that if many samples are drawn and intervals are constructed in the same way, approximately 90% of those intervals will contain the true population parameter.
- A 95% confidence level indicates that about 95 out of 100 such constructed intervals are expected to include the true parameter. This is the most widely used level in practice.

A 99% confidence level suggests a higher degree of reliability, meaning nearly 99% of the intervals will capture the true parameter.

Confidence Level and Critical Values

The confidence level determines the value of $z_{\alpha/2}$used in interval estimation:

$\boldsymbol{z_{\frac{\alpha}{2}}}$ values: 1.96 for 95% confidence

2.58 for 99% confidence

1.64 for 90% confidence

Note: As the confidence level increases, the interval becomes wider. This means higher confidence is achieved at the cost of reduced precision.

Solved Problems

1. A random sample of size 64 has a S.D. of 8. What can you say about the maximum error with 95% confidence.

Sol.

Given $\sigma = 8, n = 64, z_{\alpha/2} = 1.96$

We know that $E_{Max} = z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = 1.96\frac{8}{\sqrt{64}} = 1.96 \times 1 = 1.96$

2. A random sample of size 144 has a S.D. of 12. Find the maximum error with 99% confidence.

Sol.

Given $\sigma = 12, n = 144, z_{\alpha/2} = 2.576$

$$E_{Max} = 2.576\frac{12}{\sqrt{144}} = 2.576 \times 1 = 2.576$$

3. Assuming that $\sigma = 16$, how large a sample should be taken to assert with probability 0.95 that the sample mean will not differ from the true mean by more than 2.

Sol.

Given maximum error $E = 2$, and $\sigma = 16, z_{\alpha/2} = 1.96$ for 95% confidence

We know that

$$E_{Max} = z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \Rightarrow n = \left(\frac{z_{\alpha/2}\sigma}{E}\right)^2$$

$$n = \left(\frac{1.96 \times 16}{2}\right)^2 = (15.68)^2 = 245.86 \approx 246$$

4. Assuming that $\sigma = 25$, how large a sample should be taken to assert with probability 0.99 that the sample mean will not differ from the true mean by more than 5.

Sol.

Given $E = 5, \sigma = 25, z_{\alpha/2} = 2.576$ for 99% confidence

$$n = \left(\frac{2.576 \times 25}{5}\right)^2 = (12.88)^2 = 165.89 \approx 166$$

5. In a study of household expenses, a random sample of 100 families had a mean of Rs. 8500 and S.D. of Rs. 1200. If $\bar{x}$is used as a point estimate, with what confidence can we assert that the maximum error does not exceed Rs. 200?

Sol.

Given $n = 100$, $\bar{x} = 8500$, $\sigma = 1200$, $E_{max} = 200$

$$E_{Max} = z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \Rightarrow z_{\alpha/2} = \frac{E_{Max}\sqrt{n}}{\sigma} = \frac{200\sqrt{100}}{1200} = \frac{200 \times 10}{1200} = 1.6667$$

$z_{\alpha/2} = 1.67$

The area when $z_{\alpha/2} = 1.67$ from tables is 0.4525

$$1 - \alpha = 2 \times 0.4525 = 0.905$$

Confidence $= 90.5\%$

Hence we are 90.5% confidence that the maximum error is Rs. 200.

6. If $P = 0.4$and sample size $n = 400$, find the maximum error with 95% confidence.

Sol.

Given $P = 0.4$, $n = 400$

We have $Q = 1 - P = 0.6$and $z_{\alpha/2} = 1.96$

$$E_{Max} = 1.96\sqrt{\frac{0.4 \times 0.6}{400}} = 1.96\sqrt{\frac{0.24}{400}} = 1.96\sqrt{0.0006} = 1.96 \times 0.02449$$
$$= 0.0480$$

7. If we can assert with 95% that the maximum error is 0.04 and $P = 0.3$, find the size of the sample.

Sol.

Given $P = 0.3$, $E = 0.04$

We have $Q = 1 - P = 1 - 0.3 = 0.7$and $z_{\alpha/2} = 1.96$(for 95%)

We know that

$$E_{Max} = z_{\alpha/2}\sqrt{\frac{PQ}{n}} \Rightarrow 0.04 = 1.96\sqrt{\frac{0.3 \times 0.7}{n}}$$

$$n = \frac{0.3 \times 0.7 \times (1.96)^2}{(0.04)^2} = \frac{0.21 \times 3.8416}{0.0016} = 504.21 \approx 505$$

8. If $P = 0.5$ and sample size $n = 625$, find the maximum error with 99% confidence.

Sol.

Given $P = 0.5, n = 625$

We have $Q = 0.5$ and $z_{\alpha/2} = 2.576$

$$E_{Max} = 2.576\sqrt{\frac{0.5 \times 0.5}{625}} = 2.576\sqrt{\frac{0.25}{625}}$$

$$= 2.576 \times \sqrt{0.0004}$$

$$= 2.576 \times 0.02 = 0.0515$$

9. If we can assert with 99% that the maximum error is 0.03 and $P = 0.25$, find the size of the sample.

Sol.

Given $P = 0.25, E = 0.03$

We have $Q = 1 - P = 0.75$ and $z_{\alpha/2} = 2.576$(for 99%)

$$0.03 = 2.576\sqrt{\frac{0.25 \times 0.75}{n}}$$

$$n = \frac{0.25 \times 0.75 \times (2.576)^2}{(0.03)^2} = \frac{0.1875 \times 6.635}{0.0009} = 1382.29 \approx 1383$$

10. The mean and S.D. of a population are 850 and 120 respectively. What can one assert with 95% confidence about the maximum error if $\bar{x} = 850$ and $n = 64$. And also construct 95% confidence interval for the true mean.

Sol.

Given $\mu = 850$, $\sigma = 120$, $\bar{x} = 850$, $n = 64$, $z_{\alpha/2} = 1.96$

$$E_{Max} = z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = 1.96\left(\frac{120}{\sqrt{64}}\right) = 1.96 \times 15 = 29.4$$

Confidence interval $(\bar{x} - E_{Max}, \bar{x} + E_{Max})$

$(850 - 29.4, 850 + 29.4)$

$= (820.6, 879.4)$

Exercise: -

1. A random sample of size 100 has a S.D. of 5. What can you say about the maximum error with 95% confidence.

2. A sample of size 81 has standard deviation 9. Find the maximum error with 90% confidence.

3. Assuming that $\sigma = 20.0$, how large a random sample be taken to assert with probability 0.95 that the sample mean will not differ from the true mean by more than 3.0

4. Assuming that $\sigma = 18$, how large a sample should be taken to assert with probability 0.99 that the sample mean will not differ from the true mean by more than 4.

5. In a study of an automobile insurance a random sample of 80 body repair costs had a mean of Rs. 472.36 and the S.D. of Rs. 62.35. If $\bar{x}$ is used as a point estimate to the true average

repair costs, with what confidence we can assert that the maximum error doesn't exceed Rs. 10.

6. If we can assert with 95% that the maximum error is 0.05 and P = 0.2, find the size of the sample.

7. What is the size of the smallest sample required to estimate an unknown proportion to within a maximum error of 0.06 with at least 95% confidence.

8. The mean and S.D. of a population are 11,795 and 14054 respectively. What can one assert with 95% confidence about the maximum error if $\overline{x}$ = 11,795 and n = 50. And also construct 95% confidence interval for the true mean.

9. A sample of 10 cam shafts intended for use in gasoline engines has an average eccentricity of 1.02 and a S.D. of 0.044 inch. Assuming the data may be treated a random sample from a normal population, determine a 95% confidence interval for the actual mean eccentricity of the cam shaft?

10. A sample of 16 metal rods has an average length of 5.12 cm and a S.D. of 0.08 cm. Assuming the data may be treated as a random sample from a normal population, determine a 99% confidence interval for the true mean length of the rods.

Answers: -

1) 0.98, 2) 1.645, 3) 171, 4) 135, 5) 84.72%, 6) 246, 7) 267,

8) (7899.43,15690.57), 9) (0.993, 1.047), 10) (5.067,5.179)

Maximum Likelihood Estimation (MLE):

Suppose we have a random sample $x_1, x_2, ..., x_n$ whose assumed probability distribution depends on some unknown parameter θ. Our goal is to find good estimate of θ (population parameter) using sample and which can be done with the help of MLE.

Let $x_1, x_2, ..., x_n$ be i.i.d. random variables drawn from some probability distribution that depends on some unknown parameter θ.

The goal of MLE is to maximize the likelihood function

$$L(\theta) = f(x_1, x_2, ..., x_n \mid \theta)$$

$= f(x_1|\theta) * f(x_2|\theta) * \ldots * f(x_n|\theta)$

$L(\theta) = \prod_{i=1}^{n} f(x_i/\theta)$

$\log L(\theta) = \sum_{i=1}^{n} \log f(x_i/\theta)$

Conditions for maximization are, $\frac{dL}{d\theta} = 0$ and $\frac{d^2L}{d\theta^2} < 0$.

MLEs are Consistent, Efficient, Sufficient, MLEs may (or may not) be unbiased

MLEs are asymptotically normally distributed, asymptotically tend to have least variance.

Solved Problems

1. Consider a sample 0,1,0,0,1,0 from a binomial distribution, with the form P[X = 0] = (1-p), P[X = 1] = p. Find the maximum likelihood estimate of p.

Sol. L(p) = P[X = 0] P[X = 1] P[X = 0] P[X = 0] P[X = 1] P[X = 0]

= (1-p) p (1-p) (1-p) p (1-p)

= $(1-p)^3 p^2$.

Log L(p) = $\log[(1-p)^3 p^2.]$

= $\log[(1-p)^3] + \log[(p^{2)}.]$

= 3log(1-p) + 2logp

Conditions for maximization are $\frac{dLogL(p)}{dp} = 0$.

$$\frac{-3}{1-p} + \frac{2}{p} = 0 \Rightarrow \frac{-3p+2-2p}{p(1-p)} = 0 \Rightarrow p = 1/3$$

That is, there is 1/3 chance to observe this sample if we believe the population to be Binomial distributed.

2. An unfair coin is flipped 100 times, and 61 heads are observed. What is the MLE when nothing is previously known about the coin?

Sol. Here, the distribution follows a Binomial distribution with parameter p.
Here $n = 100$. The likelihood function (MLE) is

$$P(H = 61 \mid p) = \binom{100}{61} p^{61}(1-p)^{39}$$

For maximization

$$\frac{d}{dp} P(H = 61 \mid p) = 0$$

$$\Rightarrow \binom{100}{61} [61p^{60}(1-p)^{39} - 39p^{61}(1-p)^{38}] = 0$$

$$\Rightarrow p^{60}(1-p)^{38}(61 - 100p) = 0$$

$$\Rightarrow p = 0, \frac{61}{100}, 1$$

Thus, the likelihoods are

$$P(H = 61 \mid p = 0) = 0$$

$$P\left(H = 61 \mid p = \frac{61}{100}\right) = \binom{100}{61}\left(\frac{61}{100}\right)^{61}\left(\frac{39}{100}\right)^{39}$$

$$P(H = 61 \mid p = 1) = 0$$

Since $P\left(H = 61 \mid p = \frac{61}{100}\right)$ is maximum,

Hence $\hat{p} = \frac{61}{100}$ is the MLE.

UNIT 5

Testing of Hypothesis (Large Samples)

Hypothesis

A statement which is yet to be proved/ established or a statement on the parameter(s) of the Probability distribution to be tested. If more than one population is considered, then the statistical hypothesis gives the relation between 2 population parameters.

Test

A statistical rule which decides whether to accept or reject the null hypothesis on the basis of data.

NULL HYPOTEHSIS: This is denoted by H_0. This statement is tested for possible acceptance or rejection under the assumption that it is true.

Example 1: Suppose the average height of Soldiers is 160 cms then we setup NULL HYPOTHESIS as

H_0 : The average height of Soldiers is 160 cms i.e,

$H_0 : \mu = \mu_0$.

Example 2: Suppose we want to test the average marks between 2 teams then we setup null hypothesis as

H_0 : The average marks between 2 teams are equal. That is

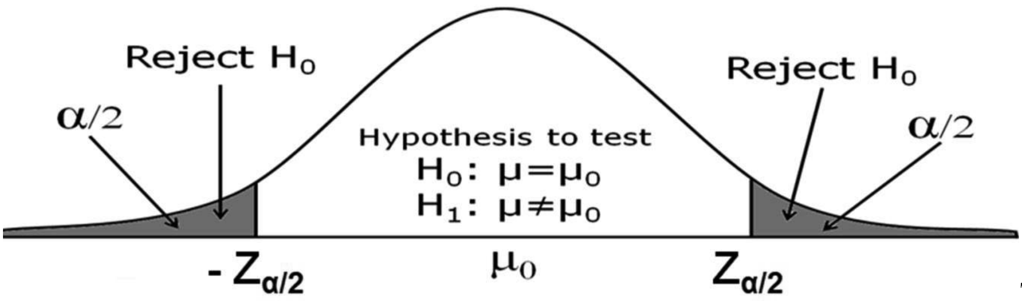
$H_0 : \mu_1 = \mu_2$.

ALTERNATIVE HYPOTHESIS: Any hypothesis which is complementary to the null hypothesis is called an alternate hypothesis and is denoted by H_1.

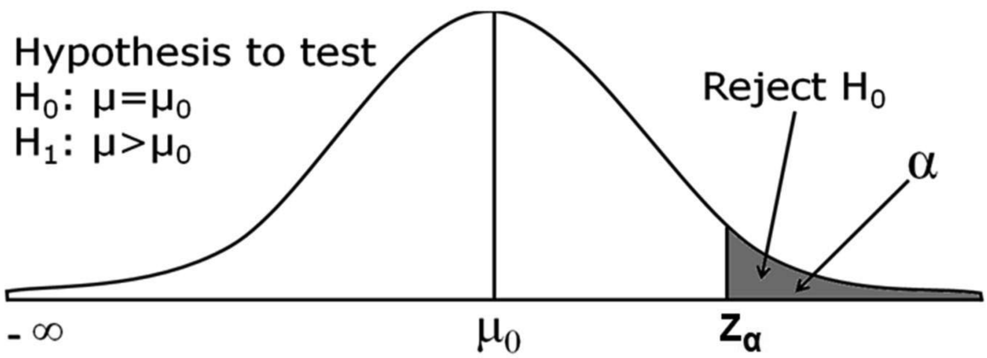
Example 1: Suppose the average height of Soldiers is 160 cms then we setup alternate hypothesis as $H_1 : \mu \neq \mu_0$.

i.e. H_1 : The average height of Soldiers is not 160 cms it is further divided into 2 statements i.e H_1: $\mu > \mu_0$. or H_1: $\mu < \mu_0$.

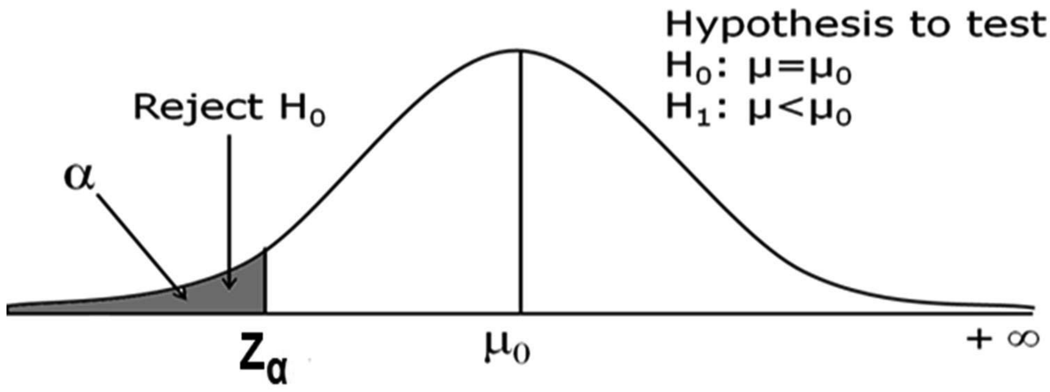
The test conducted for $\mu \neq \mu_0$ is known as **Two tailed test (TTT**) for this test the critical region lies on both ends.



The test conducted $\mu > \mu_0$ is known as **Right tailed test(RTT)** In his test the critical region lies entirely towards right side



The test conducted $\mu < \mu_0$ is known as **Left tailed test** (**LTT**) in this the critical region lies entirely towards left side



Errors in decision making

Decision is made based on the sample not on the population, this

leads to the possibility of error between the decision made and the reality.

There are two types of errors.

(i) Type I error

(ii) Type II error

(i) **Type I error**: Reject H_0 when it is true.

If the Null Hypothesis H_0 is true but it rejected by test procedure,

then the error made is called Type I error. It is denoted by α.

P(Reject H_0 when it is true) = P(Type I error) = α

(ii) **Type II error**: Accept H_0 when it is false.

If the Null Hypothesis H_0 is false but we accept H_1 by test

procedure, then the error made is called Type II error. It is denoted by β.

P(Accept H_0 when it is false) = P(Type II error) = β

Level of Significance

The level of significance denoted by α is the confidence with which we reject or accept the Null hypothesis H_0 i.e. it is the maximum possible probability with which we are willing to risk an error in rejecting H_0 when it is true.

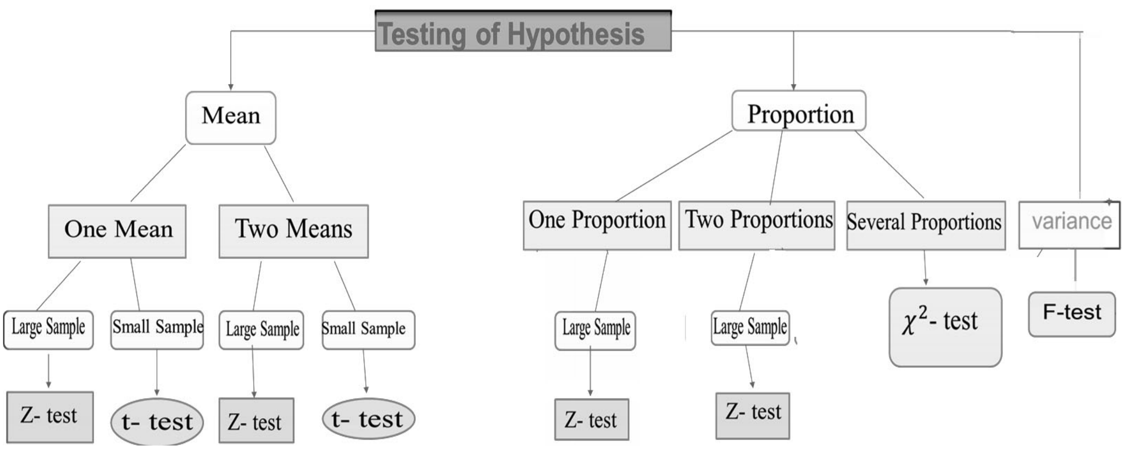
Test Statistic

Under a given hypothesis let the sampling distribution of a statistic t is approximately a normal distribution with mean E(t) then

$$z = \frac{t - E(t)}{S.E. of\ t}$$

Critical Values (Z_α) of Z

	Level of Significance		
	1%	5%	10%
Two – Tailed test	$\lvert Z_\alpha \rvert = 2.58$	$\lvert Z_\alpha \rvert = 1.96$	$\lvert Z_\alpha \rvert = 1.645$
Right – Tailed test	$Z_\alpha = 2.33$	$Z_\alpha = 1.645$	$Z_\alpha = 1.28$
Left – Tailed test	$Z_\alpha = -2.33$	$Z_\alpha = -1.645$	$Z_\alpha = -1.28$



Test of Significance for large samples

There are 4 tests of significance for large samples

(i) Test for single mean

(ii) Test for difference of means

(iii) Test for single proportion

(iv) Test for difference of proportion

I) **Procedure for testing of Hypothesis for single mean**

Let $\bar{x}$ is the mean of the random sample of size n taken from a normal population with mean μ and S.D σ.

Null Hypothesis: Set up a null hypothesis H_0: $\mu = \mu_0$

Alternative Hypothesis: Set up an Alternative Hypothesis H_1

(i) H_1: $\mu \neq \mu_0$ (two – tailed test)

(ii) H_1: $\mu > \mu_0$ (Right – tailed test)

(iii) H_1: $\mu < \mu_0$ (Left – tailed test)

Level of Significance: Select α = 5% or 1%

Test Statistic: Calculate the test statistic $Z_{cal} = \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}}$

Decision: Find the critical value Z_α or $Z_{\alpha/2}$ at the level of significance α from normal table.

If $|Z_{cal}|$ <critical value, accept H_0

If $|Z_{Cal}|$ >critical value, reject H_0 and accept H_1

1. It is claimed that sports-car owners drive on the average 16000 kms per year. A consumer firm believes that the average mileage is probably higher. To check, the consumer firm obtained information from randomly selected 40 sports-car owners that resulted in a sample mean of 16352 kms with a population standard deviation of 1348 kms. At what can be concluded about this claim at 1% level.

Sol.

Null Hypothesis: The average mileage of sports-car as claimed and the sample average mileage is the same μ = 16000.

Alternate Hypothesis: The average mileage of sports cars as claimed is higher than the sample average mileage μ >16000.

Level of Significance: α = 1%, Z_α = 2.330

Test Statistic: $z_{cal} = \frac{17352-16000}{\frac{1348}{\sqrt{40}}} = 1.650$

Decision: z_{cal} = 1.650 < 2.33, Accept H_0.

There is insufficient evidence at the 1% level of significance to conclude that sports-car owners drive more than 16000 km per year on average.

2. Suppose that we want to test on the basis of n = 35 determinations and at the 0.05 level of significance whether the thermal conductivity of a certain kind of brick is 0.340, as has been claimed. From the information gathered, we can expect that the variability of such determinations is given by σ = 0.010 and the mean of the 35 determinations is 0.343.

Sol.

Null Hypothesis: $\mu = 0.340$

Alternate Hypothesis: $\mu \neq 0.340$

Level of significance: $\alpha = 0.05$, $Z_{\frac{\alpha}{2}} = 1.96$

Test Statistic: $z_{cal} = \frac{0.343-0.340}{\frac{0.010}{\sqrt{35}}} = 1.77$

Decision: $|z_{cal}| = 1.77 < 1.96$, Accept H_0.

OR

Critical Z Range: Between -1.96 and +1.96

Since 1.78 falls within the range of -1.96 and +1.96, the test statistic does not fall into the rejection region.

At the 5%level of significance, there is insufficient evidence to reject the claim that the mean thermal conductivity of the brick is 0.340.

3. A factory claims that the average lifetime of its light bulbs is 1200 hours. A sample of 36 bulbs is tested and found to have a mean lifetime of 1220 hours with a standard deviation of 90 hours. Test the manufacturer's claim at a 5% level of significance.

Sol.

Null Hypothesis: H_0: $\mu = 1200$

Alternate Hypothesis: H_1: $\mu \neq 1200$

Level of significance: $\alpha = 0.05$, $Z_{\frac{\alpha}{2}} = 1.96$

Test Statistic: $z_{cal} = \frac{1220-1200}{\frac{90}{\sqrt{36}}} = 1.33$

Decision: $|z_{cal}| = 1.33 < 1.96$, Accept H_0.

At the 5% level of significance, there is insufficient evidence to reject the manufacturer's claim. We can conclude that the average lifetime of the light bulbs is effectively 1200 hours, and the observed difference (1220 hours) is likely due to sampling error.

Exercise

1. A sample of 64 students has a mean weight of 70kgs. Can this be regarded as a sample from a population with mean weight 56kgs and S.D. 25kgs.

2. An ambulance service claims that it takes on the average less than 10 minutes to reach its destination in emergency calls. A sample of 36 calls has a mean of 11 minutes and the variance of 16 minutes. Test the claim at 0.05 level of significance.

3. A sample of 400 items is taken from a population whose S.D. is 10. The mean of the sample is 40. Test whether the sample has come from a population with mean 38. Also calculate 95% confidence interval for the population.

II) **Procedure for testing of Hypothesis for difference of means**

Suppose $\overline{x_1}$ and $\overline{x_2}$ be the means of the two samples with sizes $n_1$and n_2 drawn from two populations having means $\mu_1$and μ_2 and S.D. $\sigma_1$and σ_2. To test whether the two-population means are equal.

Null Hypothesis: Set up a null hypothesis $H_0\, \mu_1 = \mu_2$

Alternative Hypothesis: Set appropriate alternative Hypothesis H_1

Level of Significance: Select α = 5% or 1%

Test Statistic: Calculate the test statistic

Case (i): If $\sigma_1^2 \neq \sigma_2^2 z_{cal} = \frac{\overline{x_1}-\overline{x_2}}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}$

Case (ii): If $\sigma_1^2 = \sigma_2^2 = \sigma^2 \quad z_{cal} = \frac{\overline{x_1}-\overline{x_2}}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$

Case (iii): If σ_1^2, σ_2^2 are not given instead sample variances S_1^2, S_2^2 are given then,

$$z_{cal} = \frac{\overline{x_1}-\overline{x_2}}{\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}}$$

Conclusion: If $|Z_{cal}| <$ Critical Value accept H_0

If $|Z_{Cal}| >$ Critical Value reject H_0 and accept H_1

Solved Problems

1. The manager of a courier service believes that packets delivered at the beginning of the month are heavier than those delivered at the end of month. As an experiment, he weighed a random sample of 40 packets at the beginning of the month and found that the mean weight was 5.25 kg. A randomly selected 30 packets at the end of the month had a mean weight of 4.96 kg. It was observed from the past experience that the population variances are 1.20 kg and delivered at the beginning of the month weigh more? Test at 5% level of significance the packets

Sol. Null Hypothesis: $H_0: \mu_1 = \mu_2$

Alternative Hypothesis: $H_1: \mu_1 > \mu_2$

Level of Significance: $\alpha = 5\%$, $Z\alpha = 1.645$

Test Statistic: $z_{cal} = \frac{\overline{x_1}-\overline{x_2}}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} = \frac{5.25-4.96}{\sqrt{\frac{(1.20)^2}{40}+\frac{(1.15)^2}{30}}} = \frac{5.25-4.96}{\sqrt{0.036+0.044}} = 1.025$

Decision: z_{cal} = 1.025 <z_α = 1.645, Accept H0.

At the 5%level of significance, there is insufficient evidence to conclude that packets delivered at the beginning of the month weigh more than those delivered at the end of the month.

2. The means of two large samples of sizes 1000 and 2000 members are 68.5 inches and 68.0 inches respectively. Can the samples be regarded as drawn from the same population of S.D 2.5 inches. Test at 5% level of significance.

Sol. Null Hypothesis: $\mu_1 = \mu_2$ (the samples have been drawn from the same population of S.D 2.5 inches)

Alternative Hypothesis: $\mu_1 \neq \mu_2$

Level of Significance: α = 0.05, $z_{\frac{\alpha}{2}}$ = 1.96

Test Statistic: $z_{cal} = \frac{\overline{x_1}-\overline{x_2}}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} = \frac{68..5-68}{(2.5)\sqrt{\frac{1}{1000}+\frac{1}{2000}}} = 5.16$

Decision: z_{cal} = 5.16 >$Z\frac{\alpha}{2}$ = 1.96, Reject H0.

At the 5%level of significance, the two samples cannot be regarded as drawn from the same population. There is a significant difference between the population means.

3. In a survey of buying habits, 400 shoppers are chosen at random in a super market situated in city A. Their average weekly food expenditure is Rs 275 with a S.D of Rs. 40. For 400 shoppers chosen at random in a super market situated in city B, the average weekly food expenditure is Rs.245 with a S.D of Rs. 55. Test whether the average weekly food expenditure of the two populations of shoppers is equal at 1% level of significance.

Sol Null Hypothesis: $\mu_1 = \mu_2$

Alternative Hypothesis: $\mu_1 \neq \mu_2$

Level of Significance: α = 0.01, $z_{\frac{\alpha}{2}}$ = 2.58

Test Statistic: $Z_{cal} = \frac{\overline{x_1}-\overline{x_2}}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}} = \frac{275-245}{\sqrt{\frac{(40)^2}{400}+\frac{(55)^2}{400}}} = 8.82$

Decision: $z_{cal} = 8.82 > z_{\frac{\alpha}{2}} = 2.58$, Reject H_0.

At the 1%level of significance, there is sufficient evidence to conclude that the average weekly food expenditures of shoppers in the two cities are not equal.

Exercise

1. A sample of students were drawn from two colleges and from their weights in kilograms, mean and standard deviations are calculated and shown below. Make a large sample test to test significance of the difference between the means.

	Mean	S.D.	Size of the sample
College A	55	10	400
College B	57	15	100

2. The mean yield of a particular crop from village A was 210 pounds with S.D 10 pounds per acre from a sample of 100 plots. In another village B the mean yield was 220 pounds with S.D 12 pounds from a sample of 150 plots. Assuming that the S.D of yield in the entire state was 11 pounds, test whether there is any significant difference between the mean yield of crop in the two villages.

3. Studying the flow of traffic at two busy intersections between 7pm and 9pm to determine the possible need for turn signals. It was found that on 40 week days there were on average 247.3 vehicles approaching the first intersection from the north which made a left turn, while on 30 week days there were on average 254.1 vehicles approaching the second intersection from

the north made left turns. The corresponding sample SD are 15.2 and 12. Test the significance between the difference of two means at 5% level.

III) Procedure for testing of Hypothesis for single proportion

Suppose $\bar{p}$ is proportion of the sample with size n taken from a normal population with proportion P

Null Hypothesis: set up a null hypothesis $H_0 : P = P_0$

Alternative Hypothesis: Set up an alternative hypothesis H_1

(i) $H_1 : P \neq P_0$(two – tailed test)

(ii) $H_1 : P > P_0$(Right – tailed test)

(iii) $H_1 : P < P_0$(Left – tailed test)

Level of Significance: Select α = 5% or 1%

Test Statistic: Calculate the test statistic $z_{cal} = \frac{\bar{p}-P}{\sqrt{\frac{PQ}{n}}}$

Conclusion: Find the critical value $Z_\alpha$or $Z_{\alpha/2}$at the level of significance α from the normal table.

If $|Z_{cal}|$ < critical value, accept H_0

If $|Z_{Cal}|$ >critical value, reject H_0 and accept H_1

Solved Problems

1. Experience has 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. Test the hypothesis at 0.05 level.

Sol. We have n = 400, x = 50 and p = $\frac{x}{n} = \frac{50}{400} = 0.125$

Null Hypothesis H_0 : P = 0.2

Alternative Hypothesis H_1 : P ≠ 0.2

Level of significance: α = 0.05

Test statistics $z_{cal} = \frac{\bar{p}-P}{\sqrt{\frac{PQ}{n}}} = \frac{0.125-0.2}{\sqrt{\frac{0.2*0.8}{400}}} = -3.75$

Since $|Z_{Cal}| = 3.75 > Z_{\alpha/2} = 1.96$.

At the 0.05 level of significance, we reject the null hypothesis. There is strong evidence to conclude that the proportion of top-quality products in this day's production is significantly lower than the expected 20%.

2. Natural cork in wine bottles is subject to deterioration, and as a result wine in such bottles may experience contamination. It is reported that, in a tasting of commercial wine bottles, 16 of 91 bottles were considered spoiled to some extent by cork-associated characteristics. Does this data provide strong evidence for concluding that more than 15% of all such bottles are contaminated in this way? Test the hypothesis using 0.10 level of significance.

Sol Null Hypothesis: H_0 : P = 0.15.

Alternate Hypothesis: H_1 : P > 0.15.

Level of Significance: α = 0.10

Test Statistic: $z_{cal} = \frac{\bar{p}-P}{\sqrt{\frac{PQ}{n}}} = \frac{0.1758-0.15}{\sqrt{\frac{0.15*0.85}{91}}} = 0.69$

Decision: $|Z_{Cal}| = 0.69 < Z_\alpha = 1.28$

Therefore, null hypothesis cannot be rejected.

At the 10% level of significance, there is insufficient evidence to conclude that more than 15% of all such wine bottles are contaminated by cork-associated characteristics.

3. 20 people were attacked by a virus and only 18 survived. will you reject the hypothesis that the survival rate if attacked by this virus is 85% in favour of the hypothesis that is more at 5% level.

Sol. Sample size n = 20

Number of survived people x = 18

Proportion of survived people p $= \frac{x}{n} = \frac{18}{20} = 0.9$

P = 0.85 and Q = 0.15

Null Hypothesis H_0 : P = 0.85.

Alternative Hypothesis H_1 : P > 0.85 (right one tailed test).

Level of significance: $\alpha = 0.05$

Test statistics: $z_{cal} = \frac{\bar{p}-P}{\sqrt{\frac{PQ}{n}}} = \frac{0.9-0.85}{\sqrt{\frac{0.85*0.15}{20}}} = 0.625$

Decision: $z_{cal} = 0.625 < Z_\alpha = 1.645$,

At the 5% level of significance, there is insufficient evidence to conclude that the survival rate is more than 85%. Hence, the hypothesis that the survival rate is 85% is not rejected.

We accept the Null Hypothesis H_0 .

i.e., the proportion of the survivors is 0.85.

Exercise:

1. A social worker believes that fewer than 25% of the couples in a certain area have ever used any form of fertility control. A random sample of 120 couples was contacted twenty of them said that they have used. Test the belief of the social worker at 0.05 level.

2. In a random sample of 125 soda drinkers, 68 said they prefer Thums up to Pepsi. Test the null hypothesis P = 0.5 against the alternative hypothesis P > 0.5.

3. In a sample of 1000 people in a particular country 540 prefer Brioche and the rest prefer wheatberry bread. Can we assume that both Brioche and wheatberry bread are equally popular in this country at 1% level of significance.

4. 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 0.05 level of significance.

IV) **Procedure for testing of Hypothesis for difference of proportions**

Let p_1 and p_2 be the sample proportions in two large random samples of sizes $n_1$and n_2 drawn from two populations having proportions P_1 and P_2.

Null Hypothesis: H_0 set up a null hypothesis $H_0 : \ P_1 \ = \ P_2$

Alternative Hypothesis: H_1 set appropriate alternate hypothesis

(i) $H_1 : \ P_1 \neq P_2$ (two – tailed test)

(ii) $H_1 : \ P_1 > P_2$ (right one tailed test).

(iii) $H_1 : P_1 < P_2$ (left one tailed test).

Level of Significance: α = 5% or 1%

Test Statistic: Calculate the test statistic

(a) When the population proportions P_1 and P_2 are known

$$z_{cal} = \frac{p_1 - p_2}{\sqrt{\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}}}$$

Where $Q_1 = 1 - P_1$ and $Q_2 = 1 - P_2$

(b) When the population proportions P_1 and P_2 are not known but sample proportions p_1 and p_2 are known

(i) **Method of Substitution**: In this method, sample proportions p_1 and p_2 are substituted for P_1 and P_2.

$$z_{cal} = \frac{p_1 - p_2}{\sqrt{\frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}}}$$

Where $q_1 = 1 - p_1$ and $q_2 = 1 - p_2$

(ii) **Method of Pooling**: Estimate the value of p from the two populations proportions as

$$p = \frac{n_1p_1+n_2p_2}{n_1+n_2}, q = 1 - p$$

Hence the test statistic is $$z_{cal} = \frac{p_1-p_2}{\sqrt{pq(\frac{1}{n_1}+\frac{1}{n_2})}}$$

Conclusion: If $|Z_{cal}| < Z_\alpha$ accept H_0

If $|Z_{Cal}| > Z_\alpha$ reject H_0 and accept H_1

Solved Problems

1. The article "Aspirin Use and Survival After Diagnosis of Colorectal Cancer" (J. of the Amer. Med. Assoc., 2009: 649–658) reported that of 549 study participants who regularly used aspirin after being diagnosed with colorectal cancer, there were 81 colorectal cancer-specific deaths, whereas among 730 similarly diagnosed individuals who did not subsequently use aspirin, there were 141 colorectal cancer-specific deaths. Does this data suggest that the regular use of aspirin after diagnosis will decrease the incidence rate of colorectal cancer-specific deaths? Let's test the appropriate hypotheses using a significance level of 0.05.

Sol. given $n_1 = 549, p_1 = \frac{81}{549} = 0.1475, n_2 = 730, p_2 = \frac{141}{730} = 0.1931$

p_1 be the proportion of participants who regularly used aspirin after being diagnosed with colorectal cancer.

p_2 be the proportion of similarly diagnosed individuals who did not subsequently use aspirin

Null Hypothesis $H_0 : P_1 = P_2$

Alternative Hypothesis $H_1 : P_1 < P_2$ (left one tailed test).

Level of significance: $\alpha = 0.05$

Test statistics: $z_{cal} = \frac{p_1 - p_2}{\sqrt{pq(\frac{1}{n_1}+\frac{1}{n_2})}} = \frac{0.1475-0.1931}{\sqrt{0.1736*0.8264(\frac{1}{549}+\frac{1}{730})}} = -2.41$

Where, $p = \frac{n_1p_1+n_2p_2}{n_1+n_2} = \frac{549*0.1475+730*0.1931}{549+730} = 0.1736, q = 1-p = 0.8264$

Decision: $z_{cal} = |-2.41| > Z_\alpha = 1.645$

At the 5% level of significance, the data provides sufficient evidence to conclude that regular aspirin use after diagnosis decreases the incidence rate of colorectal cancer-specific deaths.

We reject the Null Hypothesis H_0.

2. In two large cities, 30%, and 25% of the population are respectively left-handed. Is this difference between two proportions significant if samples of 1200 and 900 respectively are drawn from the two populations.

Sol. given $n_1 = 1200, p_1 = 0.3, n_2 = 900, p_2 = 0.25$

Null Hypothesis H_0 : $P_1 = P_2$. i.e. there is no significant difference.

Alternative Hypothesis H_1 : $P_1 \neq P_2$, i.e. There is a difference (two tailed test).

Level of significance $\alpha = 0.05$

Number of left-handed persons:

$$x_1 = 0.30(1200) = 360$$
$$x_2 = 0.25(900) = 225$$

Pooled proportion:

$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2}$$
$$\hat{p} = \frac{360 + 225}{1200 + 900}$$
$$\hat{p} = \frac{585}{2100}$$
$$\hat{p} = 0.2786$$
$$\hat{q} = 1 - \hat{p} = 1 - 0.2786 = 0.7214$$

Test statistics: $$z_{cal} = \frac{0.30-0.25}{\sqrt{(0.2786)(0.7214)\left(\frac{1}{1200}+\frac{1}{900}\right)}}$$

$$= \frac{0.05}{\sqrt{0.2010(0.000833 + 0.001111)}}$$

$$= \frac{0.05}{\sqrt{0.2010(0.001944)}}$$

$$= \frac{0.05}{\sqrt{0.000391}}$$

$$= \frac{0.05}{0.0198}$$

$$z_{cal} \approx 2.53$$

Decision: z_{cal} = 2.53 $>$ $z_{\frac{\alpha}{2}}$ = 1.96, We reject the Null Hypothesis H_0

The difference between the proportions of left-handed people in the two cities is significant. Hence, the two population proportions are not equal.

3. A two-wheeler manufacturing firm claims that its model A two-wheeler outsells its model B by 8%. If it is found that 42 out of a sample of 200 bike drivers prefer model A and 18 out of another sample of 100 bike drivers prefer model B, test whether the 8% difference is a valid claim.

Sol. given $n_1 = 200, p_1 = 0.21, n_2 = 100, p_2 = 0.18$

Null Hypothesis H_0 : P_1 - P_2 = 0.08 i.e., there is 8% difference in the sale of two models is a valid claim.

Alternative Hypothesis H_1 : P_1 - P_2> 0.08 (Right one tailed test).

Level of significance α = 0.05

Test statistics: $z_{cal} = \frac{(p_1-p_2)-(P_1-P_2)}{\sqrt{pq(\frac{1}{n_1}+\frac{1}{n_2})}} = \frac{0.03-0.08}{\sqrt{0.2*0.8(\frac{1}{200}+\frac{1}{100})}} = -1.02$

Where, $p = \frac{n_1p_1+n_2p_2}{n_1+n_2} = \frac{200*0.21+100*0.18}{200+100} = 0.2, q = 1-p = 0.8$

Decision: z_{cal} = $|-1.02| < Z_\alpha$ = 1.645, We accept the Null Hypothesis H_0.

.The claim that Model A outsells Model B by 8% is accepted.

Exercise

1. On the basis of their total scores, 200 candidates of a civil service examination are divided into two groups, the upper 30% and the remaining 70%. consider the first question of the examination. Among the first group ,40 had the correct answer, whereas among the second group,80 had the correct answer. On the basis of these results can one conclude that the first question is not good at discriminating ability of the type being examined here?

2. A manufacturer of electronic equipment subjects' samples of two completing brands of transistors to an accelerated performance test. If 45 of 180 transistors of the first kind and 34 of 120 transistors of the second kind fail the test, what can he conclude at the level of significance 0.05 about the difference between the corresponding sample proportions?

3. Random samples of 400 men and 600 women were asked whether they would like to have a cinema hall 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 the proposal are the same at 5% level.

UNIT 6

Testing of Hypothesis (Small Samples)

The following are some important tests for small samples

(i) Student's 't' test, (ii) F-test (iii) χ^2- test

Degrees of Freedom (d.o.f.): The number of independent variates which make up the statistic is known as the degree of freedom (d.o.f.). It is denoted by ν (the letter 'Nu' of the Greek alphabet). If n is the sample size then (d.o.f.) = ν = n-1. If there are 2 samples of sizes $\boldsymbol{n_1}$and $\boldsymbol{n_2}$ then $\nu = \boldsymbol{n_1 - n_2 - 2}$**.**

Test of Significance for small samples

There are 4 tests of significance for small samples

I. Test for single mean
II. Test for difference of means
III. Paired Sample t – Test
IV. F-Test for difference between two variances
V. F-Test for equality between several means (ANOVA)
VI. χ^2-Test for population variance
VII. Test for Goodness of fit

Student's 't' test:

I) **Procedure for testing of Hypothesis for single mean**

If $\bar{x}$ is the mean of the random sample of size n taken from a normal population with mean µ and sample Standard deviation$s$**.**

Null Hypothesis: H_0: $\mu = \mu_0$

Alternative Hypothesis: Set up an Alternative Hypothesis H_1

(i) $H_1 : \mu \neq \mu_0$ (two – tailed test)

(ii) $H_1 : \mu > \mu_0$ (Right – tailed test)

(iii) $H_1 : \mu < \mu_0$ (Left – tailed test)

Level of Significance: Select α = 5% or 1%

Test Statistic: Calculate the test statistic

(i) if sample data set is given, $t_{cal} = \frac{\bar{x}-\mu}{\frac{s}{\sqrt{n}}}$

Where $\bar{x} = \frac{\sum x}{n}$, sample standard deviation $(s) = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}$

(ii) if sample measures are given directly, $t_{cal} = \frac{\bar{x}-\mu}{\frac{s}{\sqrt{n-1}}}$

Decision: Find the critical value $t_\alpha$or $t_{\alpha/2}$at the level of significance α from the normal table.

If $|t_{cal}|$ <critical value, accept H_0

If $|t_{Cal}|$ >critical value, reject H_0 and accept H_1

Solved Problems

1. The average breaking strength of the steel rods is specified to be 18.5 thousand pounds. To test this sample of 14 rods were tested. The mean and S.D. obtained were 17.85 and 1.955 respectively. Is the result of the experiment significant?

Sol Given $\bar{x}$ = 17.85, n = 14, μ = 18.5, S.D. (s) = 1.955

Null Hypothesis H_o: μ = 18.5

Alternative Hypothesis H_1: $\mu \neq 18.5$ (two – tailed test)

Level of Significance: $\alpha = 0.05$

Test Statistic: $t_{cal} = \frac{\bar{x}-\mu}{\frac{s}{\sqrt{n-1}}} = \frac{17.85-18.5}{\frac{1.955}{\sqrt{13}}} = -1.199$

Conclusion: $t_{\frac{\alpha}{2}} = 2.16$ at 5% level of significance with n-1 = 13 d.f. from t-distribution table.

$|t_{cal}| = |-1.199| < t_{\frac{\alpha}{2}} = 2.16$

We accept H_o. Hence, there is no sufficient evidence to conclude that the average breaking strength differs from 18.5 thousand pounds.

2. It is claimed that sports-car owners drive on average mileage 18580 kms per year. A consumer firm believes that the average mileage is probably higher. To check, the consumer firm obtained information from randomly selected 10 sports-car owners that resulted in a sample mean of 17352 kms with a sample standard deviation of 2012 kms. What can be concluded about this claim at 5% level.

Sol Given $\bar{x} = 17352$, n = 10, $\mu = 18580$, S.D. (s) = 2012

Null Hypothesis H_o: $\mu = 18580$

Alternative Hypothesis H_1: $\mu > 18580$ (right – tailed test)

Level of Significance: $\alpha = 0.05$

Test Statistic: $t_{cal} = \frac{\bar{x}-\mu}{\frac{s}{\sqrt{n-1}}} = \frac{17352-18580}{\frac{2012}{\sqrt{9}}} = -1.929$

Conclusion: $t_\alpha = 1.833$ at 5% level of significance with n-1 = 9 d.o.f. from t-distribution table.

As, $t_{cal} = -1.929 < t_\alpha = 1.833$, We fail to reject H_o.

Hence, there is no evidence to support the claim that sports-car owners drive more than 18580 km per year on average.

3. In an air-pollution study performed at an experiment station, the following amounts of suspended benzene-Soluble organic matter (in micrograms per cubic meter) were obtained for eight different samples of air: 2.2, 1.8, 3.1, 2.0, 2.4, 2.0, 2.1, and 1.2. Test whether the corresponding true mean is 2.5 at 5% level.

Sol Given eight different samples of air: 2.2, 1.8, 3.1, 2.0, 2.4, 2.0, 2.1, and 1.2, n = 8, μ = 2.5.

x	$x-\bar{x}$	$(x-\bar{x})^2$
2.2	0.1	0.01
1.8	-.0.3	0.09
3.1	1	1
2.0	-0.1	0.01
2.4	0.3	0.09
2.0	-0.1	0.01
2.1	0	0
1.2	-0.9	0.81
16.8		2.02

$$s = \sqrt{\frac{2.02}{7}}$$

$$s = \sqrt{0.2886}$$

$$s \approx 0.537$$

Null Hypothesis H_o: μ = 2.5

Alternative Hypothesis H_1: μ ≠ 2.5 (two – tailed test)

Level of Significance: α = 0.05

Test Statistic

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

$$= \frac{2.1 - 2.5}{0.537/\sqrt{8}}$$

$$= \frac{-0.4}{0.537/2.828}$$

$$= \frac{-0.4}{0.190}$$

$$t \approx -2.11$$

$$\text{d.f.} = n - 1 = 8 - 1 = 7$$

At 5% level for a two-tailed test and 7 d.o.f,

$$t_{0.05} = 2.365$$

Conclusion:

$$|\, t \,| = 2.11$$

$$2.11 < 2.365$$

Therefore, we fail to reject H_0.Hence, there is no sufficient evidence to conclude that the true mean differs from 2.5 micrograms per cubic meter.

Exercise

1. A mechanic is making engine parts with axle diameters of 0.700 inch. A random sample of 10 parts shows a mean diameter of 0.742 inch with a S.D. of 0.040 inch. Test whether the work is meeting the specification at 0.05 level of significance.
2. A random sample of 10 boys had the following I.Q's : 70, 120, 110, 101, 88, 83, 95, 98, 107 and 100. Does this data support the assumption of a population mean I.Q. of 100?
3. Inspecting ceramic tiles prior to their shipment, a quality-control engineer detects 2, 3, 6, 0, 4, and 9 defects in six cartons, each containing 144 tiles. Assuming that the data can be looked upon as a random sample from a normal population, what can he assert with 99% confidence that the true average number of defectives per carton is 5?

II) Procedure for testing of Hypothesis for difference of means

Suppose $\overline{x_1}$ and $\overline{x_2}$ be the means of the two small samples with sizes n_1 and n_2 drawn from two populations having means μ_1 and μ_2 and sample S.D.'s s_1 and s_2. To test whether the two-population means are equal.

Null Hypothesis: set up a null hypothesis $H_0\ \mu_1 = \mu_2$

Alternative Hypothesis: Set appropriate alternative Hypothesis H_1

Level of Significance: Select $\alpha = 5\%$ or 1%

Test Statistic: Calculate the test statistic

Case(i): If two populations have equal variances, we use Pooled t-test

$$t_{cal} = \frac{\overline{x_1} - \overline{x_2}}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \quad \text{Where } S_p^2 = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}$$

Case(ii): If two populations have unequal variances

$$t_{cal} = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}$$

Conclusion: Find the critical value t_α or $t_{\alpha/2}$ at the level of significance α with $n_1 + n_2 - 2$ d.f. from t-distribution table.

If $|t_{cal}|$ < Critical Value, accept H_0

If $|t_{Cal}|$ > Critical Value, reject H_0 and accept H_1

Solved Problems

1. The IQ's of 16 students from one workshop showed a mean of 107 with standard deviation of 10. while the IQ's of 14 students from another workshop showed a mean of 112 with standard

deviation of 8. Is there any significant difference between the IQ's of the two groups at 5% level of significance.

Sol Given $n_1 = 16, s_1 = 10, \overline{x_1} = 107, n_2 = 14, s_2 = 8, \overline{x_2} = 112$

Null Hypothesis: H0 $\mu_1 = \mu_2$

Alternative Hypothesis: H1 $\mu_1 \neq \mu_2$

Level of Significance: α = 5%

Test Statistic: $t_{cal} = \frac{\bar{x}_1 - \bar{x}_2}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} = \frac{107-112}{9.13\sqrt{\frac{1}{16}+\frac{1}{14}}} = -1.445$

$$Where\ S_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = \frac{15 * 10^2 + 13 * 8^2}{28} = 83.28, s = 9.13$$

Decision: $t_{\frac{\alpha}{2}}$ = 2.048 at 16+14-2 = 28 d.f.

$|t_{cal}|$ = $|-1.445|$<$t_{\frac{\alpha}{2}}$ = 2.048, Accept H0.

There is no significant difference between the IQs of the two groups at the 5% level of significance.

2. Random samples of 15 and 10 were selected from two thermistors. The sample means were 315, 303 and sample standard deviations were 3.8, 4.9 respectively. Test whether there is any significant difference in the means of two thermocouples at 5% level of significance.

Sol Given $n_1 = 15, s_1 = 3.8, \overline{x_1} = 315, n_2 = 10, s_2 = 4.9, \overline{x_2} = 303$

Null Hypothesis: H0: $\mu_1 = \mu_2$

Alternative Hypothesis: H1 : $\mu_1 \neq \mu_2$

Level of Significance: α = 5%

Test Statistic

$$t = \frac{\bar{x}_1 - \bar{x}_2}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$= \frac{315 - 303}{4.26\sqrt{\frac{1}{15} + \frac{1}{10}}} = \frac{12}{4.26(0.408)} = \frac{12}{1.738}$$

$$t \approx 6.90$$

Where

$$S_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

$$S_p^2 = \frac{14(3.8)^2 + 9(4.9)^2}{15 + 10 - 2}$$

$$= \frac{14(14.44) + 9(24.01)}{23}$$

$$= \frac{202.16 + 216.09}{23}$$

$$S_p^2 \approx 18.18$$

$$S_p = \sqrt{18.18} \approx 4.26$$

Decision

At 5% level of significance for a two-tailed test and d.o.f. $= n_1 + n_2 - 2 = 15 + 10 - 2 = 23, t_{0.025} = 2.069$

$$|\, t \,| = 6.90 > t_{0.025} = 2.069$$

Therefore, we reject H_0. There is a **significant difference** between the mean values of the two thermocouples at the 5% level of significance.

3. Two athletes X and Y were tested according to the time (in seconds) to run a particular track with the following results.

Athletes X	28	30	32	33	33	29	34
Athletes Y	29	30	30	24	27	29	

Test whether the two athletes have the same running capacity.

Sol Given $n_1 = 7, n_2 = 6$

$\bar{x} = \sum\frac{x}{n_1} = 31.286, \bar{y} = \sum\frac{y}{n_2} = 28.16$

x	$x-\bar{x}$	$(x-\bar{x})^2$	y	$y-\bar{y}$	$(y-\bar{y})^2$
28	-3.286	10.8	29	0.84	0.7056
30	-1.286	1.6538	30	1.84	3.3856
32	0.714	0.51	30	1.84	3.3856
33	1.714	2.94	24	-4.16	17.3056
33	1.714	2.94	27	-1.16	1.3456
29	-2.286	5.226	29	0.84	0.7056
34	2.714	7.366			
219		31.4358	169		26.8336

S.D. $(S_p) = \sqrt{\frac{\sum(x-\bar{x})^2+\sum(y-\bar{y})^2}{n_1+n_2-2}} = \sqrt{\frac{31.4358+26.836}{11}} = \sqrt{5.23} = 2.286$

Null Hypothesis: H0 $\mu_1 = \mu_2$ i.e. the two athletes have the same running capacity.

Alternative Hypothesis: $H_1\ \mu_1 \neq \mu_2$

Level of Significance: α = 5%, $t_{\frac{\alpha}{2}}$ = 2.201 at 7+6-2 = 11 d.f.

Test Statistic: $t_{cal} = t = \frac{\bar{x}_1 - \bar{x}_2}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} = \frac{31.286-28.16}{2.286\sqrt{\frac{1}{7}+\frac{1}{6}}} = 2.443$

Decision: t_{cal} = 2.443 >$t_{\frac{\alpha}{2}}$ = 2.201, Reject H_0. There is a significant difference between the running capacities of the two athletes at the 5% level of significance.

Exercise

1. The means of two random samples of sizes 9 and 7 are 196.42 and 198.82 respectively. The sum of the squares of the deviations from the mean are 26.94 and 18.73 respectively. Can the sample be considered to have been drawn from the same normal population.

2. Samples of two types of light bulbs were tested for length of life and following data were obtained

	Type I	Type II
Sample size	n_1 = 8	n_2 = 7
Sample mean hours	$\bar{x}$ = 1234	$\bar{y}$ = 1036 hrs
Sample S.D.	s_1 = 36 hrs	s_2 = 40 hrs

Is the difference in the means sufficient to warrant that type I is superior to type II regarding length of life.

III) **PAIRED – SAMPLE t – TEST**

This test is used to test whether any significance difference before and after some experiment or training etc.

Null Hypothesis: $H_0: \mu_d = 0$

Alternative Hypothesis: $H_1: \mu_d \neq 0$(two – tailed test)

$H_1: \mu_d > 0$ (Right tailed test)

$H_1: \mu_d < 0$ (Left tailed test)

where $\mu_d$is the mean difference.d = After - Before. Or d = Before-After

Level of Significance: Select α = 5% or 1%

Test Statistic: For paired t-test,

$$t = \frac{\bar{d}}{s_d/\sqrt{n}} \text{ where } \bar{d} = \frac{\sum d_i}{n} \text{and } s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(d_i - \bar{d})^2$$

Conclusion: Find the critical value $t_\alpha$or $t_\alpha/2$ at the level of significance α with $n-1$ d.f. from t-distribution table.

If $|t_{cal}|$ < Critical Value, accept H_0

If $|t_{Cal}|$ > Critical Value, reject H_0 and accept H_1

1. The blood Pressure of 5 men before and after intake of a certain syrup are given below

Before	110	120	125	132	125
After	120	118	125	136	121

Test whether there is significant change in Blood Pressure at 1% level of significance.

Sol$N.H.\,H_0: \mu_d = 0$

A.H. $H_1: \mu_d \neq 0$ (a two-tailed test)

where μ_d is the mean difference.

$Take\ d$ = After − Before Person	d = After − Before
1	10
2	-2
3	0
4	4
5	-4

$$\sum d = 10 - 2 + 0 + 4 - 4 = 8$$

$$\bar{d} = \frac{8}{5} = 1.6$$

d	$d - \bar{d}$	$(d - \bar{d})^2$
10	8.4	70.56
-2	-3.6	12.96

d	$d - \bar{d}$	$(d - \bar{d})^2$
0	-1.6	2.56
4	2.4	5.76
-4	-5.6	31.36

$$\sum(d - \bar{d})^2 = 123.20$$

$$s_d = \sqrt{\frac{123.20}{5-1}}$$

$$= \sqrt{30.8}$$

$$s_d \approx 5.55$$

Test Statistic: For paired t-test,

$$t = \frac{\bar{d}}{s_d/\sqrt{n}}$$

$$t = \frac{1.6}{5.55/\sqrt{5}}$$

$$= \frac{1.6}{5.55/2.236}$$

$$= \frac{1.6}{2.48}$$

$$t \approx 0.645$$

$$\text{d.o.f.} = n - 1 = 5 - 1 = 4$$

At 1% level of significance for a two-tailed test and 4 d.f., $t_{0.01} = 4.604$

Decision

$$| t | = 0.645$$

$$0.645 < 4.604$$

Therefore, fail to reject H_0. There is no significant change in blood pressure after intake of the syrup at the 1% level of significance.

IV) SNEDECOR'S F – TEST OF SIGNIFICANCE

F – test is used to test whether there is any significance difference between the variances. Let two independent random samples of sizes n_1 and n_2 be drawn from two normal populations. To test the hypothesis that the two population variances σ_1^2 and σ_2^2 are equal.

Null Hypothesis: H0 $\sigma_1^2 = \sigma_2^2$

Alternative Hypothesis: H1 $\sigma_1^2 \neq \sigma_2^2$

Level of Significance: Select α = 5% or 1%

Test Statistic: $F_{cal} = \frac{S_1^2}{S_2^2}$ if $S_1^2 > S_2^2$ or $F_{cal} = \frac{S_2^2}{S_1^2}$ if $S_2^2 > S_1^2$

$$S_1^2 = \frac{n_1 s_1^2}{n_1 - 1} = \frac{\sum(x-\bar{x})^2}{n_1 - 1}, \; S_2^2 = \frac{n_2 s_2^2}{n_2 - 1} = \frac{\sum(y-\bar{y})^2}{n_2 - 1}$$

Where s_1^2 and s_2^2 are the variances of the two samples.

Conclusion: Find the critical value F_α at the level of significance α with $\nu_1 = n_1 - 1$ and $\nu_2 = n_2 - 1$ d.f. from F-distribution table.

If $|F_{cal}| < F_\alpha$ accept H0

If $|F_{cal}| > F_\alpha$, reject H0 and accept H1

Solved Problems

1. In one sample of 8 observations from a normal population, the sum of the squares of deviations of the sample values from the sample mean is 84.4 and in another sample of 10 observations it was 102.6. Test at 5% level whether the populations have the same variance.

Sol Given $n_1 = 8$, $n_2 = 10$, $\sum(x-\bar{x})^2 = 84.4$, $\sum(x-\bar{x})^2 = 102.6$,

$$S_1^2 = \frac{\sum(x-\bar{x})^2}{n_1-1} = \frac{84.4}{7} = 12.057,\ S_2^2 = \frac{\sum(y-\bar{y})^2}{n_2-1} = \frac{102.6}{9} = 11.4$$

Null Hypothesis: $H_0\ \sigma_1^2 = \sigma_2^2$

Alternative Hypothesis: $H_1\ \sigma_1^2 \neq \sigma_2^2$

Level of Significance: Select $\alpha = 5\%$

Test Statistic: $F_{cal} = \frac{S_1^2}{S_2^2} = \frac{12.057}{11.4} = 1.057$ since, $S_1^2 > S_2^2$

Conclusion: $F_{0.05}(7,9) = 3.29$ from the F-distribution table.

$F_{cal} = 1.057 < F_{0.05} = 3.29$ accept H_0

i.e. the populations have the same variance.

2. Two random samples reveal the following results.

Sample	Size	Sample Mean	Sum of Squares of deviations from the Mean
1	10	15	90
2	12	14	108

Test whether the samples came from the same normal population with equal variances.

Sol Given n_1 = 10, n_2 = 12, $\sum(x-\bar{x})^2 = 90, \sum(y-\bar{y})^2 = 108$

$$S_1^2 = \frac{\sum(x-\bar{x})^2}{n_1-1} = \frac{90}{9} = 10,$$

$$S_2^2 = \frac{\sum(y-\bar{y})^2}{n_2-1} = \frac{108}{11} = 9.81$$

To belong to the same normal population, both population parameters must match:

$$\mu_1 = \mu_2 \text{ and } \sigma_1^2 = \sigma_2^2$$

This involves two tests:

1. Test equality of variances → F-test

2. Test equality of means → t-test (using pooled variance)

<u>Test for Equality of Variances</u>

Null Hypothesis: $H_0\ \sigma_1^2 = \sigma_2^2$

Alternative Hypothesis: $H_1\ \sigma_1^2 \neq \sigma_2^2$

Level of Significance: Select α = 5%

Test Statistic: $F_{cal} = \frac{S_1^2}{S_2^2} = \frac{10}{9.81} = 1.019$ since, $S_1^2 > S_2^2$

Conclusion: $F_{0.05}(9,11) = 3.10$ from the F-distribution table.

$F_{cal} = 1.019 < F_{0.05} = 3.1$, accept H_0

<u>Test for Equality of Means</u>

Now use pooled variance t-test.

Null Hypothesis H_0: $\mu_1 = \mu_2$

Alternative Hypothesis H_1: $\mu_1 \neq \mu_2$

$$S_p^2 = \frac{90+108}{10+12-2}$$

$$= 9.9$$

$$S_p = \sqrt{9.9} = 3.146$$

Test Statistic: $t = \dfrac{\bar{x}_1 - \bar{x}_2}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} = \dfrac{15-14}{3.146\sqrt{\frac{1}{10}+\frac{1}{12}}}$

$$= \frac{1}{3.146\sqrt{0.1 + 0.0833}}$$

$$t \approx 0.74$$

d.o.f. $n_1 + n_2 - 2 = 20$

Critical value at 5% level: $t_{0.05,20} = 2.086$

Since, $0.74 < 2.086$

we accept H_0.

The difference between the sample means is not significant and the variances are also equal. Hence, the two samples may be regarded as having come from the same normal population with equal variances.

(V) Test for Equality of several means (ANOVA)

G_1	G_2	G_3	. . .	G_k
X_{11}	X_{21}	X_{31}	. . .	X_{k1}
X_{12}	X_{22}	X_{32}	. . .	X_{k2}
X_{13}	X_{23}	X_{33}	. . .	X_{k3}
⋮	⋮	⋮	⋮	⋮
X_{1n1}	X_{2n2}	X_{3n1}	. . .	X_{kn1}
C_1	C_2	C_3	. . .	C_k

$n_1 + n_2 + n_3 + \ldots + n_k = n$

$C_1 + C_2 + C_3 + \ldots + C_k = G$

Calculation of sum of squares

1. Correction factor (CF): $\frac{G^2}{n}$
2. Total sum of squares (TSS): $\sum_i \sum_j x_{ij}^2 - CF$
3. Between group sum of squares (GSS): $\sum_{i=1}^{k} \frac{C_i^2}{n_i} - CF$
4. Within group sum of squares (WSS): TSS – GSS

ANOVA Table

Source of variation	df	Sum of squares	Mean sum of squares	F-ratio
Between groups	k - 1	GSS	$MGSS = \frac{GSS}{k-1}$	$F = \frac{MGSS}{MWSS}$
Within groups	n - k	WSS	$MWSS = \frac{WSS}{n-k}$	
Total	n - 1	TSS	$F \approx$ F−distribution with k−1 and n−k df	

MGSS- Mean group sum of squares

MWSS- Mean within group sum of squares

Null Hypothesis: H_0 $\mu_1 = \mu_2 = \ldots = \mu_k$

Alternative Hypothesis: H_1 $\mu_1 \neq \mu_2 \neq \ldots \neq \mu_k$ (Not all the means are equal)

Level of Significance: Select α = 5% or 1%

Test Statistic: $F = \frac{MGSS}{MWSS}$

Conclusion: If $F_{cal} > \approx F_{(k-1,n-k)}$ with k-1,n-k d.o.f. null Hypothesis H_0 is rejected. Otherwise H_0 is accepted.

Solved Problem

1. Setup an ANOVA table for the following per acre production data for 3 varieties of wheat each grown on 4 plots and state if the variety differences are significant.

Plot of Land	Variety of Wheat		
	A	B	C
1	6	5	5
2	7	5	4
3	3	3	3
4	8	7	4

Sol. Null Hypothesis: $H_0\ \mu_1 = \mu_2 = \mu_3$

Alternative Hypothesis: $H_1\ \mu_1 \neq \mu_2 \neq \mu_3$ (Not all the means are equal)

Level of Significance: Select $\alpha = 5\%$

Calculation of sum of squares

1. Correction factor (CF): $\frac{G^2}{n} = \frac{60^2}{12} = 300$
2. Total sum of squares (TSS): $\sum_i \sum_j x_{ij}^2 - CF = 32$
3. Between group sum of squares (GSS): $\sum_{i=1}^{k} \frac{C_i^2}{n_i} - CF$
$$= \frac{24^2}{4} + \frac{20^2}{4} + \frac{16^2}{4} - 300 = 8$$
4. Within group sum of squares (WSS): TSS – GSS = 32-8 = 24

ANOVA Table

Source of Variation	d.o.f	Sum of Squares	Mean Sum of Squares	F-ratio
Between Groups	3-1 = 2	8	$MGSS = \frac{8}{2} = 4$	$F = \frac{4}{2.67} = 1.498$
Within Groups	12-3 = 9	24	$MWSS = \frac{24}{9} = 2.67$	
Total	11	32		

Test Statistic: $F = \frac{MGSS}{MWSS} = \frac{4}{2.67} = 1.498$

Conclusion: If $F_{cal} = 1.498 < F_{(2,9)} = 4.26$ with (2,9)d.o.f.

Therefore, Null Hypothesis H_0 is accepted.

2. A test was given to 5 students taken at random from 5th standard of 3 schools of a city. The individual scores are given below. Test if there is a significant difference between schools.

Students	Schools		
	School-1	School-2	School-3
Student-1	9	7	6
Student-2	7	4	5
Student-3	6	5	6
Student-4	5	4	7
Student-5	8	5	6

Sol. Null Hypothesis: $H_0\ \mu_1 = \mu_2 = \mu_3$

Alternative Hypothesis: $H_1\ \mu_1 \neq \mu_2 \neq \mu_3$(Not all the means are equal)

Level of Significance: Select $\alpha = 5\%$

Calculation of sum of squares

1. Correction factor (CF): $\frac{G^2}{n} = \frac{90^2}{15} = 540$
2. Total sum of squares (TSS): $\sum_i \sum_j x_{ij}^2 - CF = 568 - 540 = 28$
3. Between group sum of squares (GSS):

$$\text{GSS} = \sum_{i=1}^{k} \frac{C_i^{\,2}}{n_i} - \text{CF}$$

$$= \frac{35^2}{5} + \frac{25^2}{5} + \frac{30^2}{5} - 540$$

$$= 10$$

4. Within group sum of squares (WSS): TSS – GSS = 28-10 = 18

ANOVA Table

Source of Variation	d.o.f	Sum of Squares	Mean Sum of Squares	F-ratio
Between Groups	3-1 = 2	10	$MGSS = \frac{10}{2} = 5$	$F = \frac{5}{1.5} = 3.33$
Within Groups	15-3 = 12	18	$MWSS = \frac{18}{12} = 1.5$	
Total	14	28		

Test Statistic: $F = \frac{MGSS}{MWSS} = \frac{5}{1.5} = 3.33$

Conclusion: If $F_{cal} = 3.33 < F_{(2,12)} = 3.89$ with (2,12)d.o.f.

Therefore, Null Hypothesis H_0 is accepted.

Exercise

1.The following data give the yields on 12 plots of land in three samples under 3 varieties of fertilizers.

A	B	C
25	20	24
22	17	26
24	16	30
21	19	20

Test at 5% level of significance whether there is any difference between the yield of the varieties.

F (2, 9) at 5% level is 4.26

2. Three drying formulas for curing glue are studied

Formula A	13	10	8	11	8	
Formula B	13	11	14	14		
Formula C	4	1	3	4	2	4

Test at 5% level of significance whether is any difference in the mean curing time of glue?

F(0.05; 2, 12) = 3.885

(VI) χ^2-TEST FOR POPULATION VARIANCE

To test whether the population variance σ^2 is equal to σ_0^2 using sample data.

Null Hypothesis: $H_0\ \sigma^2 = \sigma_0^2$

Alternative Hypothesis: $H_1\ \sigma^2 \neq \sigma_0^2$

Level of Significance: Select α = 5% or 1%

Test Statistic: ${\chi^2}_{cal} = \frac{(n-1)s^2}{\sigma^2} = \frac{\sum(x-\bar{x})^2}{\sigma^2}$

Conclusion: If $\chi^2{}_{cal} > \chi^2{}_{\alpha}$ with n-1 d.f. the Null Hypothesis H_0 is rejected. Otherwise H_0 is accepted.

Solved Problems

1. Plastic sheeting produced by a machine must be periodically monitored for possible fluctuations in thickness. Uncontrollable variation in the viscosity of the liquid mold produces some variation in thickness. Based on experience with a great many samples, when the machine is working well, an observation on thickness has a normal distribution with standard deviation σ = 1.2 mm. Samples of 20 thickness measurements are collected regularly. A value of the sample standard deviation exceeding 1.4mm signals concern about the product. Find whether σ = 1.2 mm, at 5% level.

Sol Null Hypothesis: $H_0\ \sigma^2 = (1.2)^2$

Alternative Hypothesis: $H_1\ \sigma^2 > (1.2)^2$

Level of Significance: Select α = 5%

Test Statistic: $\chi^2{}_{cal} = \frac{(n-1)s^2}{\sigma^2} = \frac{19*(1.4)^2}{(1.2)^2} = 25.86$

Conclusion: If $\chi^2{}_{cal} = 25.86 < \chi^2{}_{\alpha} = 30.1$ with n-1 = 19 d.o.f. the null Hypothesis H_0 is accepted. There is no significant evidence at the 5% level to conclude that the population standard deviation exceeds 1.2 mm.

2. An optical firm purchases glass to be ground into lenses, and it is known from past experience that the variance of the refractive index of this kind of glass is $1.26 * 10^{-4}$. As it is important that the various pieces of glass have nearly the same index of refraction, the firm receives a shipment of sample variance of 20 pieces selected at random to be $2.00 * 10^{-4}$. Assuming that the sample values may be looked upon as a random sample from a normal population, can we reject $\sigma^2 = 1.26 * 10^{-4}$ at 5% level.

Sol Null Hypothesis: $H_0\ \sigma^2 = 1.26 * 10^{-4}$

Alternative Hypothesis: $H_1\ \sigma^2 \neq 1.26 * 10^{-4}$

Level of Significance: Select α = 5%

Test Statistic: ${\chi^2}_{cal} = \frac{(n-1)s^2}{\sigma^2} = \frac{19*(2.00*10^{-4})}{1.26*10^{-4}} = 30.2$

Conclusion: If ${\chi^2}_{cal} = 30.2 > {\chi^2}_{\alpha} = 30.144$ with n-1 = 19 d.o.f.

Null Hypothesis H_0 is rejected. There is sufficient evidence to conclude that the variance of the refractive index in this shipment is significantly higher than the expected $1.26\ X10^{-4}$

3. A manufacturer claims that any of his list of items cannot have variance 1 cm². A sample of 25 items has a variance of 1.2 cm². Test whether the claim of the manufacturer is correct.

Sol Null Hypothesis: $H_0\ \sigma^2 = 1$

Alternative Hypothesis: $H_1\ \sigma^2 \neq 1$

Level of Significance: Select α = 5%

Test Statistic: ${\chi^2}_{cal} = \frac{(n-1)s^2}{\sigma^2} = \frac{24*(1.2)}{1} = 28.8$

Conclusion: If ${\chi^2}_{cal} = 28.8 < {\chi^2}_{\alpha} = 36.41$ with n-1 = 24 d.f.

Null Hypothesis H_0 is accepted. Hence the claim of the manufacturer is correct.

Exercise

1. Test the claim that the variance of a normal population is 21.3 is rejected at 5% level of significance if the variance of a random sample of size 15 is 39.74.

(VII) a) Chi-Square Test As A Test For Goodness Of Fit:

The test which is used to find how good a fit is there between observed and expected frequencies is called test for goodness of fit. χ^2 distribution is used to test how well a fit is there between observed frequency and expected (theoretical) frequency distribution.

$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$ with υ d.o.f.

Null Hypothesis: H_0 The discrepancy between observed and expected frequencies is not significant. i.e. It is a good fit

Alternative Hypothesis: H_1 The discrepancy between observed and expected frequencies is significant. i.e. It is not a good fit

Level of Significance: Select α = 5% or 1%

Test Statistic: ${\chi^2}_{cal} = \sum \frac{(O_i - E_i)^2}{E_i}$

Conclusion: If ${\chi^2}_{cal} > {\chi^2}_{\alpha}$ the Null Hypothesis H_0 is rejected. Otherwise H_0 is accepted.

Note:

1. If $\chi^2 = 0, O_i$agrees exactly with E_i.
2. If $\chi^2 > 0$, and

   b) If χ^2 is small, then $O_i and\ E_i$ are closer to each other indicating a good fit.
   c) If χ^2 is large, then $O_i are\ away\ from\ E_i$ indicating a poor fit.

Conditions for applying χ^2-Test.

1. N≥ 50
2. $4 \leq k \leq 16$
3. $\sum O_i = \sum E_i$
4. O_i or E_i ≥ 10, otherwise combine the neighbouring frequencies so that O_i or E_i ≥ 10

Solved Problems

1. The number of promotional calls per week of a certain product are as follows: 12, 8, 20, 2,14,10,15, 6, 9, 4. Are these frequencies in agreement with the belief that promotional calls were the same during this 10 week period at 5% level.

Sol. Null Hypothesis: H_0 The promotional calls were the same during the 10 week period.

Alternative Hypothesis: H_1 The promotional calls were not the same during the 10 week period.

Level of Significance: Select α = 5%

Test Statistic: Expected frequency of promotional calls each week $E_i = \frac{100}{10} = 10$

Observed Frequency (O_i)	Expected Frequency (E_i)	$(O_i - E_i)$	$\frac{(O_i - E_i)^2}{E_i}$
8	10	2	0.4
20	10	-2	0.4
2	10	10	10.0
14	10	-8	6.4
10	10	4	1.6
15	10	0	0.0
6	10	5	2.5
9	10	-4	1.6
4	10	-1	0.1
	10	-6	3.6
100	100		26.6

$$\chi^2{}_{cal} = \sum \frac{(O_i - E_i)^2}{E_i} = 26.6$$

Conclusion: $\chi^2{}_{\alpha}$ = 16.9 at 5% level of significance at n-1 = 10-1 = 9 d.o.f.

$$\chi^2_{cal} = 26.6 > \chi^2_{\alpha} = 16.9, \text{Reject } H_0$$

i.e. the number of promotional calls were not the same during the 10 week period

2. A sample analysis of examination results of 500 students was made. It was found that 220 students had failed, 170 had secured a third class, 90 were placed in second class and 20 got a first class. Do these figures commensurate with the general examination result which is in the ratio of 4 : 3 : 2 : 1 for the various categories respectively.

Sol Null Hypothesis: H_o The observed results commensurate with the general examination results.

Alternative Hypothesis: H_1 The observed results are not commensurate with the general examination results.

Level of Significance: $\alpha = 5\%$

Test Statistic: Expected frequencies are in the ratio of 4 : 3 : 2 : 1

Total frequency = 500,

we divide the total frequency 500 in the ratio 4 : 3 : 2 : 1, we get the expected frequencies as 200, 150, 100, 50

Observed Frequency (O_i)	Expected Frequency (E_i)	$(O_i - E_i)$	$\frac{(O_i - E_i)^2}{E_i}$
220	200	20	2.00
170	150	20	2.667
90	100	-10	1.000
20	50	-30	18.00
500	500		23.667

$$\chi^2_{cal} = \sum \frac{(O_i - E_i)^2}{E_i} = 23.667$$

Conclusion: $\chi^2{}_{\alpha} = 7.81$ at 5% level of significance at n-1 = 4-1 = 3 d.o.f.

$\chi^2{}_{cal} = 23.667 > \chi^2{}_{\alpha} = 7.81$, Reject H_0

i.e. The observed results are not commensurate with the general examination results.

3. A die is thrown 264 times with the following results. Show that the die is biased

No. appeared on the die	1	2	3	4	5	6
Frequency	40	32	28	58	54	52

Ans. $\chi^2{}_{cal} = 17.64 > \chi^2{}_{\alpha} = 11.07$, Reject H0. The observed frequencies differ significantly from the expected frequencies of a fair die. Hence, the die is biased.

Exercise

1. A certain type of flashlight is Sold with the four batteries included. A random sample of 150 flashlights is obtained, and the number of defective batteries in each is determined, resulting in the following data:

Number Defective	0	1	2	3	4
Frequency	26	51	47	16	10

Test the null hypothesis that distribution of X is binomial where X is the number of defective batteries in a randomly selected flash light (χ^2 table value at 4 df = 9.488 at 5% level)

2. The following is the distribution of the hourly number of trucks arriving at a company's warehouse.

Trucks arriving per hour	0	1	2	3	4	5	6	7	8
Frequency	52	151	130	102	45	12	5	1	2

Fit a Poisson distribution and test for goodness of fit at 5% level of significance at (χ^2 table value at 4 df = 9.488 at 5% level).

c) To Test the Independence of Attributes

To test whether the two different attributes are independent or not we use χ^2 test.

Consider a 2 X 2 contingence table

Categorical Variable 1	Categorical variable 2		Total
	Present	**Absent**	
Present	O_1 E_1	O_2 E_2	r_1
Absent	O_3 E_3	O_4 E_4	r_2
Total	c_1	c_2	n

Calculation of expected frequencies

$$E_1 = \frac{r_1 c_1}{n} \qquad E_2 = \frac{r_1 c_2}{n}$$

$$E_3 = \frac{r_2 c_1}{n} \qquad E_4 = \frac{r_2 c_2}{n}$$

Null Hypothesis: H_o The two attributes are independent

Alternative Hypothesis: H_1 The two attributes are not independent

Level of Significance: α = 5% or 1%

Test Statistic: $\chi^2{}_{cal} = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$

where k = r x c is the total number of cells in the r x c contingency table, r = total no. of rows

c = total no. of columns.

Conclusion: If $\chi^2{}_{cal} > \chi^2{}_{\alpha}$ with (r-1Xc-1) d.o.f. the Null Hypothesis H_0 is rejected. Otherwise H_0 is accepted

Solved Problems

1. The opinion of a random sample of 500 employees about 3 Health Insurance Plans are shown below. Are they independent of job classification?

Job Classification	Health Insurance Plan			Total
Salaried workers	166	86	68	320
Hourly workers	84	64	32	180
Total	250	150	100	500

Sol Null Hypothesis: H_o The two attributes are independent

Alternative Hypothesis: H_1 The two attributes are not independent

Level of Significance: $\alpha = 5\%$

Test Statistic: $\chi^2{}_{cal} = \sum_{i=1}^{6} \frac{(O_i - E_i)^2}{E_i} = 3.921$

$E_1 = \frac{r_1 c_1}{n} = \frac{320*250}{500} = 160.00$

$E_2 = \frac{r_1 c_2}{n} = \frac{320*150}{500} = 96.00$

$E_3 = \frac{r_1 c_3}{n} = \frac{320*100}{500} = 64.00$

$E_4 = \frac{r_2 c_1}{n} = \frac{180*250}{500} = 90.00$

$E_5 = \frac{r_2 c_2}{n} = \frac{180*150}{500} = 54.60$

$E_6 = \frac{r_2 c_3}{n} = \frac{180*100}{500} = 36.00$

S No	O_i	E_i	$O_i - E_i$	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
1	166	160.00	6	36	0.225
2	86	96.00	-10.00	100.00	1.04
3	68	64.00	4.00	16.00	0.25
4	84	90.00	-6.00	36.00	0.40
5	64	54.60	9.40	88.36	1.62
6	32	36.00	-4.00	16.00	0.44
Total	180	180	Chi-square value $(\chi^2{}_{cal})$		3.921

Conclusion: If $\chi^2{}_{cal} < \chi^2{}_{\alpha} = 5.992$ with (2-1X3-1) = 2 d.o.f.

The Null Hypothesis H_0 is accepted. Therefore, Job classification and health insurance plans are independent.

Exercise

1. Examine whether the nature of the area is related to the voting preference in any election.

Area	Votes for		Total
	A	B	
Rural	620	380	1000
Urban	550	450	1000
Total	1170	830	2000

UNIT 7

CORRELATION AND REGRESSION

CORRELATION

1 Introduction to Correlation

Correlation: Correlation is a statistical concept that describes the relationship between two variables. It indicates how changes in one variable are associated with changes in another variable. The relationship between variables may exist in different forms. Variables may move in the same direction (both increase or decrease), in opposite directions, or may show no clear pattern.

When two variables vary together, either in the same direction or in opposite directions, they are said to be correlated. correlation is useful in fields such as economics, business, engineering, and science for analysis, prediction, and decision-making.

2 Types of Correlation

(i) Correlation based on Direction

Positive Correlation: When two variables move in the same direction, that is, both increase or decrease together, the correlation is called positive correlation.

Eg: Positive Correlation: Increase in height and weight of a person

Negative Correlation: When one variable increases while the other decreases, the correlation is called negative correlation.

Eg: Negative Correlation: Increase in price leads to decrease in demand.

(ii) Correlation based on Number of Variables

Simple Correlation: It studies the relationship between two variables only.

Eg: Relationship between income and expenditure.

Multiple Correlation: It examines the relationship among three or more variables simultaneously.

Eg: Relationship among price, demand, and supply of a commodity

(iii) Correlation based on Form of Relationship

Linear Correlation: The change in one variable is proportional to the change in another, which results in a straight-line relationship.

Eg: Distance travelled at constant speed over time.

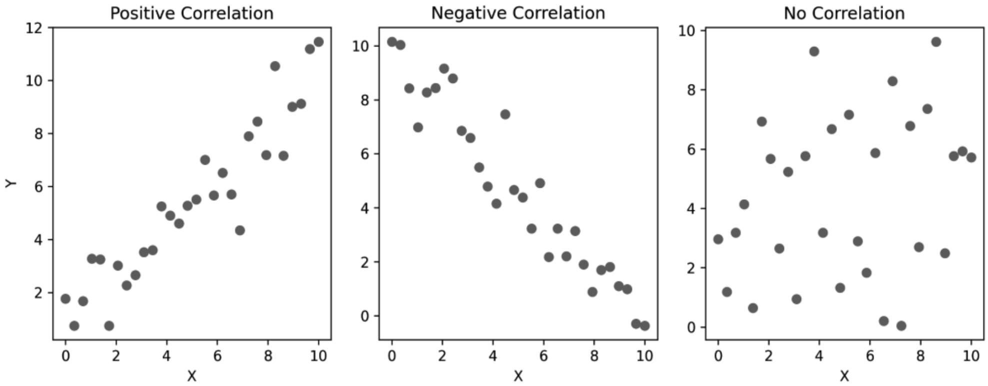
Non-linear Correlation: The rate of change is not constant, and the relationship forms a curve.

Eg: Age and physical strength of a person.

3 Methods of Studying Correlation

Correlation can be studied using two main approaches: graphical methods and mathematical methods.

(a) Graphical Methods: Scatter diagram is a graphical method in which by assuming one variable on the X-axis and the other one on the Y-axis, values of two variables are plotted as points on a graph. The pattern formed by these points indicates the type of correlation.



- Interpretation of scatter plots
  - ➢ If the points move upward from left to right, it indicates positive correlation.
  - ➢ If the points move downward from left to right, it indicates negative correlation.
  - ➢ If the points are scattered without any clear pattern, there is no correlation.

(b) Mathematical Methods

- Karl Pearson's coefficient method
- Spearman's rank correlation method

4 Karl Pearson's Correlation Coefficient

It is a widely used statistical measure that indicates the degree and direction of linear relationship between two variables. It is denoted by *r*.

Karl Pearson's coefficient can be calculated using three different formulas

1. $r = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2 \sum(y-\bar{y})^2}}$

2. $r = \frac{n\sum xy-(\sum x)(\sum y)}{\sqrt{[n\sum x^2-(\sum x)^2][n\sum y^2-(\sum y)^2]}}$

3. $r = \frac{Cov(x,y)}{\sigma_x \sigma_y}$

Properties of Karl Pearson's Correlation Coefficient

The value of r always lies between −1 and +1 and indicates both the strength and direction of the relationship.

- $r = +1$, it indicates perfect positive correlation
- $r = -1$, it indicates perfect negative correlation
- $r = 0$, it indicates no correlation
- $0 < r < 1$, it indicates positive correlation
- $-1 < r < 0$, it indicates negative correlation

Problems

1. Calculate coefficient of correlation from the following data:

X	5	7	6	9	8	10	4
Y	7	9	8	10	9	12	5

Sol We use the formula $r = \frac{\text{Cov}(X,Y)}{\sigma(X)\sigma(Y)}$

Computation of coefficient of correlation

X	Y	X^2	Y^2	XY
5	7	25	49	35
7	9	49	81	63
6	8	36	64	48
9	10	81	100	90
8	9	64	81	72
10	12	100	144	120
4	5	16	25	20

$$\Sigma X = 49, \Sigma Y = 60, \Sigma X^2 = 371$$
$$\Sigma Y^2 = 544, \Sigma XY = 448, N = 7$$

$$r = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{[N\Sigma X^2 - (\Sigma X)^2][N\Sigma Y^2 - (\Sigma Y)^2]}}$$

$$r = \frac{7(448) - (49 \times 60)}{\sqrt{[7(371) - (49)^2][7(544) - (60)^2]}}$$
$$= \frac{3136 - 2940}{\sqrt{[2597 - 2401][3808 - 3600]}}$$
$$= \frac{196}{\sqrt{196 \times 208}}$$
$$= \frac{196}{\sqrt{40768}} = \frac{196}{201.91}$$
$$r = 0.97$$

2: Find if there is any significant correlation between X and Y given below:

X	48	50	52	54	56	58	60
Y	100	104	108	110	114	118	120

Sol

Coefficient of correlation

$$r = \frac{\Sigma XY}{\sqrt{\Sigma X^2 \cdot \Sigma Y^2}}$$

Computation of coefficient of correlation

(Take assumed means: $A = 54$, $B = 110$)

X	**X = x – 54**	**X²**	**Y**	**Y = y – 110**	**Y²**	**XY**
48	-6	36	100	-10	100	60
50	-4	16	104	-6	36	24

52	-2	4	108	-2	4	4
54	0	0	110	0	0	0
56	2	4	114	4	16	8
58	4	16	118	8	64	32
60	6	36	120	10	100	60
Total	**0**	**112**		**0**	**320**	**188**

$$r = \frac{188}{\sqrt{112 \times 320}}$$
$$= \frac{188}{\sqrt{35840}} = \frac{188}{189.32}$$
$$r = 0.99$$

Exercise:-

1. Calculate coefficient of correlation from the following data:

X	12	9	8	10	11	13	7
Y	14	8	6	9	11	12	3

2.Find if there is any significant correlation between the heights and weights given below:

Height (in inches)	57	59	62	63	64	65	55	58	57
Weight (in lbs)	113	117	126	126	130	129	111	116	112

Answers:

1. $r = +0.95$,   2. r = +0.98

5 Spearman's Rank Correlation

Spearman's rank correlation measures the degree of relationship between two variables based on their ranks rather than actual values. It is useful when data is given in order or when exact numerical values are not available.

Formula: $\rho = \mathbf{1} - \frac{\mathbf{6\sum d^2}}{\boldsymbol{n(n^2-1)}}$ where

d = difference between ranks of corresponding items

n = number of observations

Properties

- ρ = +1 : Perfect positive correlation
- ρ = −1 : Perfect negative correlation
- ρ = 0 : No correlation

Formula (With Ties):

When two or more items have the same value, they are assigned the average of the ranks they would have occupied. After assigning average ranks to tied items, the next rank is taken as the position following the last rank used in the tie. For example, if two items share ranks 2 and 3 (each assigned 2.5), the next item will be given rank 4. The formula is given by

$$r_s = 1 - \frac{6\left[\sum d^2 + \frac{1}{12}(m_1^3 - m_1) + \frac{1}{12}(m_2^3 - m_2) + \cdots\right]}{n(n^2-1)}$$

Where m =number of tied ranks

Solved Problems:

1. Following are the ranks obtained by 10 students in two subjects, Statistics and Mathematics. To what extent the knowledge of the students in two subjects is related?

Statistics	1	2	3	4	5	6	7	8	9	10
Mathematics	3	1	4	2	6	5	8	7	10	9

Sol

Rank in Statistics (x)	Rank in Mathematics (y)	D = (x − y)	D²
1	3	-2	4
2	1	+1	1
3	4	-1	1
4	2	+2	4
5	6	-1	1
6	5	+1	1
7	8	-1	1
8	7	+1	1
9	10	-1	1
10	9	+1	1
Total			16

$$D^2 = 16$$

$$\rho = 1 - \frac{6\Sigma D^2}{N(N^2-1)}$$

$$\rho = 1 - \frac{6 \times 16}{10(10^2-1)}$$

$$= 1 - \frac{96}{10(100-1)} = 1 - \frac{96}{990}$$

$$= 1 - 0.097$$

$$\rho = 0.90$$

2. The following data represent marks obtained by 8 students in Machine Learning (ML) and Data Science (DS).

Student	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**
ML	78	65	82	90	72	60	85	68
DS	75	70	80	88	68	65	84	72

Find Spearman's rank correlation coefficient.

Sol

Assign ranks (Highest mark = Rank 1)

ML	Rank x	DS	Rank y	$D = x - y$	D^2
78	4	75	4	0	0
65	7	70	6	1	1
82	3	80	3	0	0
90	1	88	1	0	0
72	5	68	7	-2	4
60	8	65	8	0	0
85	2	84	2	0	0
68	6	72	5	1	1
					$\sum \boldsymbol{D^2} = \mathbf{6}$

$$\rho = 1 - \frac{6\sum D^2}{n(n^2-1)}$$

$$= 1 - \frac{6(6)}{8(8^2-1)}$$

$$= 1 - \frac{36}{8(63)}$$

$$= 1 - \frac{36}{504}$$

$$= 1 - 0.0714$$

$$\boxed{\rho = 0.93}$$

Hence, there is a high positive correlation between the marks.

3. The following data show marks obtained by students in two subjects.

Student	A	B	C	D	E	F	G	H
Artificial Intelligence (AI)	70	85	85	60	75	90	70	80
Data Mining (DM)	68	88	84	62	78	92	68	81

Find Spearman's rank correlation coefficient.

Sol

Ranks with ties are assigned average ranks.

AI	Rank x	DM	Rank y	D	D^2
70	6.5	68	6.5	0	0
85	2.5	88	2	0.5	0.25
85	2.5	84	3	-0.5	0.25
60	8	62	8	0	0
75	5	78	5	0	0
90	1	92	1	0	0
70	6.5	68	6.5	0	0
80	4	81	4	0	0
					$\sum \boldsymbol{D^2} = \mathbf{0.5}$

Tie correction:

In Economics, 85 occurs twice and 70 occurs twice.

$$\frac{1}{12}(2^3 - 2) + \frac{1}{12}(2^3 - 2) = \frac{6}{12} + \frac{6}{12} = 1$$

In Commerce, 68 occurs twice.

$$\frac{1}{12}(2^3 - 2) = \frac{6}{12} = 0.5$$

Total correction:

$$1 + 0.5 = 1.5$$

Formula with ties:

$$\rho = 1 - \frac{6(\sum D^2 + \text{Correction})}{n(n^2 - 1)}$$

$$= 1 - \frac{6(0.5 + 1.5)}{8(8^2 - 1)}$$

$$= 1 - \frac{12}{504}$$

$$= 1 - 0.0238$$

$$\boxed{\rho = 0.976}$$

Hence, there is a very high positive correlation.

Exercise:

1. The following are the ranks obtained by 9 students in Physics and Chemistry. Find Spearman's rank correlation coefficient.

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

2. The following data represent marks obtained by students in English and History. Find Spearman's rank correlation coefficient.

Student	A	B	C	D	E	F	G
English	65	72	80	60	75	68	85
History	70	74	82	58	78	65	88

3. The following data show scores obtained in two tests. Find Spearman's rank correlation coefficient.

Test 1	45	60	60	75	80	80	90
Test 2	50	65	68	78	78	85	92

Answers:

1. 0.97, **2.** 0.96, **3.** 0.95

REGRESSION

1 Introduction to Regression

Regression is a statistical method used to estimate or predict the value of one variable based on the value of another related variable. It helps in studying how one variable depends on another by providing a functional relationship. It enables us to determine the expected value of one variable when the other is known.

Correlation indicates the strength and direction of a relationship but does not provide a predictive equation. Regression, on the other hand, gives an equation that can be used for estimation and prediction of values.

Regression is widely used in practical situations such as estimating demand based on price, predicting sales from advertising expenditure, or estimating weight from height. It helps in making decisions and forecasting future outcomes.

2 Regression Lines

A regression line is a straight line that represents the average relationship between two variables. It is used to estimate the value of one variable based on the known value of another variable.

For a given set of data, there are always two regression lines—one for Y on X and another for X on Y—and both pass through the point of means $(\bar{x}, \bar{y})$.

Regression of Y on X: The regression line of Y on X is used to estimate the value of Y for a given value of X. In this case, X is treated as the independent variable and Y as the dependent variable.

Regression of X on Y: The regression line of X on Y is used to estimate the value of X for a given value of Y. Here, Y is considered as the independent variable and X as the dependent variable.

3 Method of Least Squares

In this method, the best line is obtained by minimizing the sum of the squares of the deviations (errors) between the actual values and the values predicted by the regression line. This ensures that the fitted line represents the data as closely as possible.

In regression analysis, the relationship between variables may take different mathematical forms depending on the nature of data.

1. A straight line represents a constant rate of change

2. A parabola represents a relationship where the rate of change is not constant but varies with the variable.

3. An exponential curve shows rapid growth or decay

4. A power curve represents a proportional relationship between variables

5. A logarithmic curve reflects a slow rate of increase.

The selection of a suitable curve depends on how the data behaves and the pattern it follows.

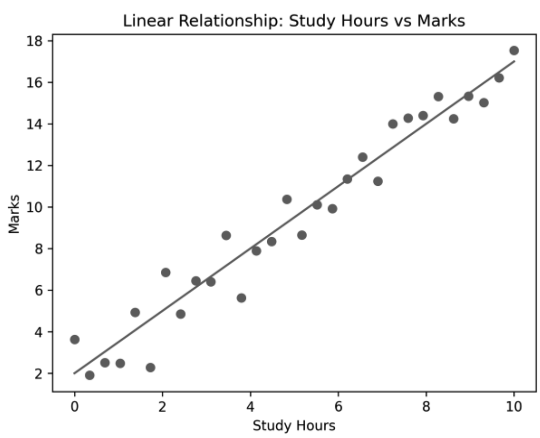
The method of least squares is used to find the best fitting curve by minimizing the sum of squares of errors between observed and estimated values.

Let the observed values be (x, y) and the estimated value be Y.
Condition: $\sum (y - Y)^2$ is minimum

3.1. Straight Line (Linear Relationship)

A straight line represents a relationship where the change in one variable is constant with respect to another. This means for every equal increase in X, Y increases or decreases by a fixed amount.

Example: Study hours and marks as study time increases, marks tend to increase steadily.



Assume the equation: $Y = a + bX$

Error: $e = y - (a + bX)$

Minimize: $\sum [y - (a + bX)]^2$

Normal Equations

$\sum y = n\,a + b \sum x$

$\sum xy = a \sum x + b \sum x^2$

Solve these equations to obtain values of a and b.

Solved Problems:

1. A company records advertising expenditure x in thousand rupees and sales y in lakh rupees. Fit a straight line and estimate sales when the expenditure is Rs. 7,000.

X	2	4	6	8	10
Y	15	22	29	35	42

Sol

x	y	x^2	xy
2	15	4	30
4	22	16	88
6	29	36	174
8	35	64	280
10	42	100	420

From the table, $\Sigma x = 30$, $\Sigma y = 143$, $\Sigma x^2 = 220$, $\Sigma xy = 992$, and $n = 5$.

The normal equations for the straight line $y = a + bx$ are:

$$\Sigma y = na + b\Sigma x$$

$$\Sigma xy = a\Sigma x + b\Sigma x^2$$

Substituting the values in the normal equations gives:

$$143 = 5a + 30b$$

$$992 = 30a + 220b$$

Solving the two simultaneous equations, we get $a = 8.7$ and $b = 3.3167$.

Therefore, the fitted straight line is:

$$y = 8.7 + 3.3167x$$

Estimation: For $x = 7$, $y = 8.7 + 3.3167(7) = 31.9169$. Hence, the expected sales are approximately 31.92 lakh rupees.

2. Fit a straight line $y = a + bx$by the method of least squares for the following data:

x	2	4	6	8	10
y	8	13	18	22	29

Sol

x	Y	x^2	xy
2	8	4	16

x	Y	x^2	xy
4	13	16	52
6	18	36	108
8	22	64	176
10	29	100	290
Σx=30	Σy=90	Σx^2= 220	Σxy =642

$$n = 5,\ \Sigma x = 30,\ \Sigma y = 90$$

Normal equations:

$$\Sigma y = na + b\Sigma x$$
$$90 = 5a + 30b$$
$$\Sigma xy = a\Sigma x + b\Sigma x^2$$
$$642 = 30a + 220b$$

Solving, $a = 3.0, b = 2.5$

Hence the required straight line is $y = 3 + 2.5x$

3. By the method of least squares fit a straight line for the following data:

X	5	10	15	20	25	30
Y	42.8	39.4	35.9	32.7	29.1	25.8

Sol

Assume the straight line is $y = a + bx$

x	**y**	**x²**	**xy**
5	42.8	25	214.0
10	39.4	100	394.0
15	35.9	225	538.5
20	32.7	400	654.0
25	29.1	625	727.5
30	25.8	900	774.0
Σ	205.7	2275	3302.0

Also, $\Sigma x = 105, n = 6$

The normal equations are

$$\Sigma y = na + b\Sigma x$$
$$205.7 = 6a + 105b$$

and

$$\Sigma xy = a\Sigma x + b\Sigma x^2$$
$$3302 = 105a + 2275b$$

Solving, $a = 46.18, b = -0.68$

Hence the required straight line is $y = 46.18 - 0.68x$

Exercise:

1. By the method of least squares, find the straight line that best fits the following data:

X	1	2	3	4	5
Y	14	27	40	55	68

2. Fit a straight line of the form $y = a + bx$for the following data:

X	0	5	10	15	20	25
Y	12	15	17	22	24	30

3. Fit a straight line to the following data by the method of least squares:

X	0	1	2	3	4
Y	1	1.8	3.3	4.5	6.3

4. By the method of least squares fit a straight line for the following data:

X	10	15	20	25	30	35
Y	35.3	32.4	29.2	26.1	23.2	20.5

5. The following data show study hours x per day and marks y obtained by students. Fit a straight line and estimate the marks for a student who studies 5 hours per day.

X	1	2	3	4	6
Y	3	4	5	55	6

6. Fit a straight line for the data:

X	2	4	6	8	10
Y	5	7	9	8	11

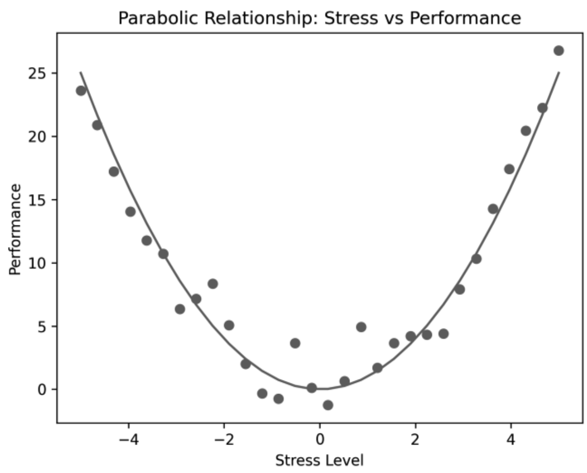
Answers:

1. $a = 0, b = 13.6, y = 13.6x$ **2.** $a = 11.2862, b = 0.6971, y = 11.2862 + 0.6971x$ **3.** $a = 0.72, b = 1.33, y = 0.72 + 1.33x$ **4.** $a = 44.18, b = -0.68, \quad y = 44.18 - 0.68x$ **5.** y = 6.7162x + 28.1081, y = 61.69 at x = 5 **6.** y = 4.1 + 0.6x

3.2. Parabola (Quadratic Relationship)

A parabolic curve represents a relationship where the rate of change varies. The variable may increase up to a certain point and then decrease, forming a curved pattern.

Example: Stress vs performance — performance increases with moderate stress but decreases when stress becomes too high.



Assume the equation: $Y = a + bX + cX^2$

Minimize: $\sum [y - (a + bX + cX^2)]^2$

Normal Equations

$$\sum y = n\,a + b\sum x + c\sum x^2$$
$$\sum xy = a\sum x + b\sum x^2 + c\sum x^3$$
$$\sum x^2y = a\sum x^2 + b\sum x^3 + c\sum x^4$$

Solve these equations to obtain a, b and c.

Solved Problems :

1. The monthly demand of a product is observed for five consecutive months. Fit a second-degree polynomial and estimate the trend.

X	1	2	3	4	5
Y	5	8	14	23	35

Sol consider a second-degree polynomial

$$y = a + bx + cx^2$$

For a second-degree polynomial, the normal equations are

$$\Sigma y = na + b\Sigma x + c\Sigma x^2$$

$$\Sigma xy = a\Sigma x + b\Sigma x^2 + c\Sigma x^3$$

$$\Sigma x^2 y = a\Sigma x^2 + b\Sigma x^3 + c\Sigma x^4$$

Computation table:

x	y	x^2	x^3	x^4	xy	x^2y
1	5	1	1	1	5	5
2	8	4	8	16	16	32
3	14	9	27	81	42	126
4	23	16	64	256	92	368
5	35	25	125	625	175	875

n = 5

$$\Sigma x = 15, \Sigma y = 85$$
$$\Sigma x^2 = 55, \Sigma x^3 = 225, \Sigma x^4 = 979$$
$$\Sigma xy = 330, \Sigma x^2 y = 1406$$

Substituting these values in the normal equations:

$$85 = 5a + 15b + 55c$$

$$330 = 15a + 55b + 225c$$

$$1406 = 55a + 225b + 979c$$

Solving the above simultaneous equations:

$a = 5$, $b = -1.5$, $c = 1.5$

Final fitted polynomial**:** $y = 5 - 1.5x + 1.5x^2$

2. Fit a second degree polynomial $y = a + bx + cx^2$ by the method of least squares for the following data:

x	1	2	3	4	5
y	4	9	17	28	42

Sol Assume the parabola $y = a + bx + cx^2$

x	y	x^2	x^3	x^4	xy	x^2y
1	4	1	1	1	4	4
2	9	4	8	16	18	36
3	22	9	27	81	66	198
4	35	16	64	256	140	560
5	56	25	125	625	280	1400
Σ	126	55	225	979	508	2198

Also, $n = 5, \sum x = 15$

The normal equations are

$$\sum y = na + b\sum x + c\sum x^2$$
$$\sum xy = a\sum x + b\sum x^2 + c\sum x^3$$
$$\sum x^2 y = a\sum x^2 + b\sum x^3 + c\sum x^4$$

Substituting,

$$126 = 5a + 15b + 55c$$
$$508 = 15a + 55b + 225c$$
$$2198 = 55a + 225b + 979c$$

Solving these equations, $a = 3, b = -1, c = 2$

Hence the required parabola is $y = 3 - x + 2x^2$

3. Fit a parabola of the form $y = a + bx + cx^2$ by the method of least squares for the following data:

x	0	2	4	6	8
y	2.1	6.0	16.1	31.8	53.9

Sol Given equation of parabola is $y = a + bx + cx^2$

x	y	x^2	x^3	x^4	xy	x^2y
0	2.1	0	0	0	0	0
2	6.0	4	8	16	12.0	24.0
4	16.1	16	64	256	64.4	257.6
6	31.8	36	216	1296	190.8	1144.8
8	53.9	64	512	4096	431.2	3449.6
$\sum x = 20$	$\sum y = 109.9$	$\sum x^2 = 120$	$\sum x^3 = 800$	$\sum x^4 = 5664$	$\sum xy = 698.4$	$\sum x^2 y = 4876$

Also, $n = 5, \sum x = 20$

The normal equations are

$$109.9 = 5a + 20b + 120c$$

$$698.4 = 20a + 120b + 800c$$

$$4876 = 120a + 800b + 5664c$$

Solving, $a = 2.06, b = 0.83, c = 0.72$

Therefore, $y = 2.06 + 0.83x + 0.72x^2$

Exercise:

1) Fit a second degree polynomial to the following data by the method of least squares:

x	0	1	2	3	4
y	1	1.8	3.3	4.5	6.3

2) By the method of least squares fit a parabola of the form $y = a + bx + cx^2$ for the following data:

x	2	4	6	8	10
y	3.07	12.85	31.47	57.38	91.29

3) By the method of least squares fit a parabola of the form $y = a + bx + cx^2$ for the following data:

x	10	15	20	25	30	35
y	35.3	32.4	29.2	26.1	23.2	20.5

4) Fit a parabola to the following data by the method of least squares:

x	0	1	2	3	4
y	1	1.8	3.3	4.5	6.3

5) The fuel consumption of a vehicle changes with speed. Fit a quadratic relation between speed index x and fuel consumption y.

x	1	2	3	4	5
y	12	15	20	27	36

Answers:

1) $a = 0.962, b = 0.844, c = 0.1214,\ y = 0.962 + 0.844x + 0.1214x^2$

2) $a = 0.696, b = -0.8551, c = 0.992,\ y = 0.696 - 0.8551x + 0.992x^2$

3) $a = 41.9257, b = -0.6690, c = 0.001571,\ y = 36.8471 - 0.6837x + 0.01532x^2$

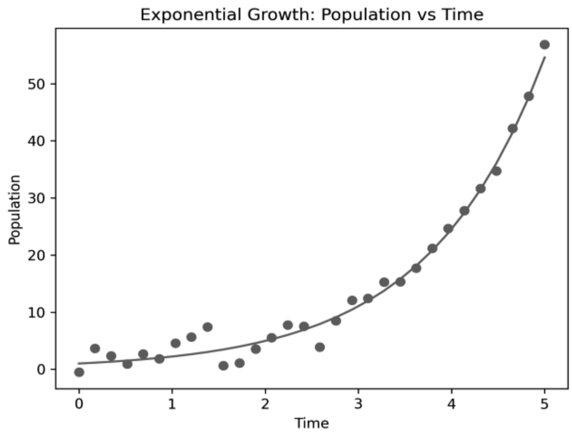
4) $a = 0.962, b = 0.844, c = 0.1214,\ y = 0.962 + 0.844x + 0.1214x^2$

5) $y = 11 + x^2$

3.3. Exponential Curve

An exponential curve shows rapid growth or decay, where the value increases slowly at first and then very quickly.

Example: Population growth over time — population increases slowly initially and then grows rapidly.



Assume the equation: $Y = a\, e^{bX}$

Take logarithm on both sides: $\log Y = \log a + bX$

Let $Y' = \log Y$ and $A = \log a$

Then $Y' = A + bX$

Normal Equations

$\sum Y' = n\,A + b \sum X$

$\sum XY' = A \sum X + b \sum X^2$

Solve to get A and b, then a = antilog A

Problem

1. The population of bacteria in a culture is recorded at different times $x$hours. Fit an exponential curve. Data: $\{x = 1,2,3,4,5\}$ and $\{y = 12,20,33,54,90\}$

Sol

Taking natural logarithm on both sides, $\ln\ y = \ln\ a + bx$

Let $Y = \ln\ y, A = \ln\ a$

Then the equation becomes $Y = A + bx$

which is a straight line.

Table

X	y	Y = ln y	x^2	xY
1	12	2.4849	1	2.4849
2	20	2.9957	4	5.9914
3	33	3.4965	9	10.4895
4	54	3.9890	16	15.9560
5	90	4.4998	25	22.4990
15	209	17.4659	55	57.4208

Number of observations: $n = 5$

For the line $Y = A + bx$

the normal equations are

$$\sum Y = nA + b\sum x$$

$$\sum xY = A\sum x + b\sum x^2$$

$$17.4659 = 5A + 15b$$
$$57.4208 = 15A + 55b$$

Solving the equations we get $b = 0.5023$ and $A = 1.9863$

Since $A = \ln\ a$

$a = e^{1.9863} \approx 7.29$

Therefore the fitted exponential curve is $y = 7.29e^{0.502x}$

2. Determine the constants $a$and b by the method of least squares such that $y = ae^{bx}$ fits the following data:

x	0	1	2	3	4
y	2.00	4.45	9.91	22.05	49.06

Sol Given curve $y = ae^{bx}$

Taking natural logarithms, $\ln\ y = \ln\ a + bx$

Let $Y = \ln\ y, A = \ln\ a$

Then $Y = A + bx$

X	Y	$Y = \ln\ y$	xY	x^2
0	2.00	0.6931	0	0
1	4.45	1.4929	1.4929	1
2	9.91	2.2935	4.5870	4
3	22.05	3.0933	9.2799	9

X	Y	$Y = \ln\, y$	xY	x^2
4	49.06	3.8931	15.5724	16
Σx=10		Σ Y=11.4659	Σ xY=30.9322	Σx^2= 30

Also, $n = 5, \sum x = 10$

Normal equations:

$$\sum Y = nA + b\sum x$$
$$11.4659 = 5A + 10b$$
$$\sum xY = A\sum x + b\sum x^2$$
$$30.9322 = 10A + 30b$$

Solving, $A = 0.6931, b = 0.80$

Hence $a = e^{0.6931} = 2$

Therefore $y = 2e^{0.8x}$

3. Find the curve of best fit of the type $y = ae^{bx}$ for the following data by the method of least squares:

X	1	2	3	4	5
Y	1.65	2.72	4.48	7.39	12.18

Sol Given curve $y = ae^{bx}$

Taking natural logarithms, $\ln\, y = \ln\, a + bx$

Let $Y = \ln\ y, A = \ln\ a$

Then $Y = A + bx$

X	Y	$Y = \ln\ y$	xY	x^2
1	1.65	0.5008	0.5008	1
2	2.72	1.0006	2.0012	4
3	4.48	1.4996	4.4988	9
4	7.39	2.0001	8.0004	16
5	12.18	2.5006	12.5030	25
Σx=15		ΣY=7.5017	ΣxY=27.5042	Σx^2=55

Also, $n = 5, \sum x = 15$

Normal equations:

$$7.5017 = 5A + 15b$$
$$27.5042 = 15A + 55b$$

Solving, $A \approx 0.001, b \approx 0.50$

Hence $a = e^{0.001} \approx 1$

Therefore $y = e^{0.5x}$

Exercise:

1) Determine the constants $a$and b by the method of least squares such that $y = ae^{bx}$ for the following data:

X	2	4	6	8	10
Y	4.077	11.084	30.128	81.897	222.62

2) Find the curve of best fit of the type $y = ae^{bx}$ for the following data by the method of least squares:

X	1	5	7	9	12
Y	10	15	12	15	21

3) Using the method of least squares, find the constants a and b such that $y = ae^{bx}$ fits 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

4) Fit the curve $y = ae^{bx}$ to the following data:

X	0	1	2	3	4	5	6	7	8
Y	20	30	52	77	135	211	326	550	1052

5) The number of bacteria in a culture was recorded at different times. Fit an exponential curve $\mathbf{y = ae^{bx}}$.

x	**0**	**1**	**2**	**3**	**4**
y	50	82	135	220	360

Answers

1) $A = 0.405, a = 1.499, b = 0.5, y = 1.499e^{0.5x}$

2) $A = 2.2484, a = 9.4725, b = 0.059, y = 9.4725e^{0.059x}$

3) $A = -2.2832,\ a = 0.10195, b = 2.9963,\ y = 0.10195e^{2.9963x}$

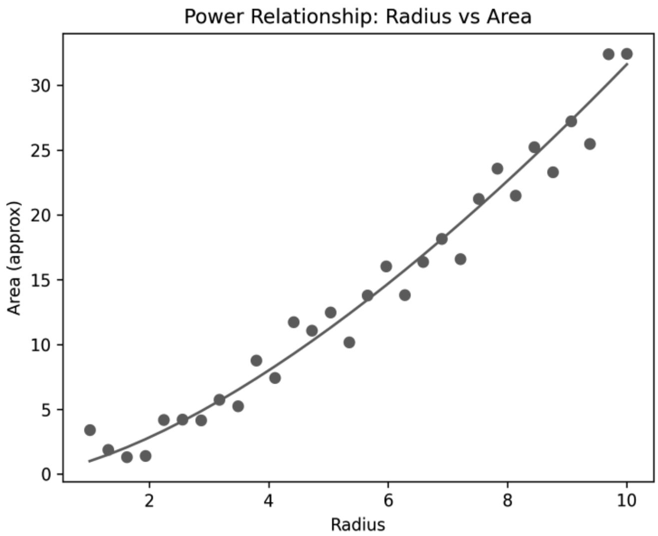
4) $A = 2.93895,\ a = 18.8960, b = 0.4876,\ y = 18.8960e^{0.4876x}$

5) $y\ =\ 50.0857.\,\mathrm{e}0.4935\mathrm{x}$

3.4. Power Curve

A power curve represents a relationship where one variable changes as a power of another. The growth is not constant but follows a specific proportional pattern.

Example: Radius and area — area increases as a power of radius.



Assume the equation: $Y = a\,X^b$

Take logarithm: $\log Y = \log a + b \log X$

Let $Y' = \log Y$ and $X' = \log X$

Then $Y' = A + b\,X'$

Normal Equations

$\sum Y' = n\,A + b \sum X'$

$\sum X'Y' = A \sum X' + b \sum (X'^2)$

Solve to get A and b, then a = antilog A

Solved Problem:

1.The production rate y units/hour of a machine is recorded for different operating speeds x rpm. Fit a power curve. Data: x = 100, 150, 200, 250, 300 and y = 35, 50, 68, 82, 96.

Sol A power curve has the form $y = ax^b$

Taking logarithm on both sides,

Using natural logarithm, $\ln\ y = \ln\ a + b\ln\ x$

Let $Y = \ln\ y, X = \ln\ x, A = \ln\ a$

Then the equation becomes $Y = A + bX$

which is a straight line.

Calculation Table

x	y	$X = \ln\ x$	$Y = \ln\ y$	X^2	XY
100	35	4.6052	3.5553	21.2079	16.3736
150	50	5.0106	3.9120	25.1061	19.5985
200	68	5.2983	4.2195	28.0720	22.3565
250	82	5.5215	4.4067	30.4870	24.3307
300	96	5.7038	4.5643	32.5333	26.0365
		26.1394	20.6578	137.4063	108.6958

For the straight line $Y = A + bX$

the normal equations are

$$\sum Y = nA + b\sum X$$
$$\sum XY = A\sum X + b\sum X^2$$

Substituting values,

$$20.6578 = 5A + 26.1394b$$
$$108.6958 = 26.1394A + 137.4063b$$

Solving the Equations

From the first equation,

$$A = \frac{20.6578 - 26.1394b}{5}$$

Substituting into the second equation and Solving, $b = 0.9294$

Substituting back, $A = -0.7271$

Since

$$A = \ln\, a$$
$$a = e^{-0.7271} \approx 0.4833$$

Hence the required power curve is

$$y = 0.4833\, x^{0.9294}$$

2. Fit an exponential curve of the form $y = ab^x$ to the following data.

X	1	2	3	4	5
Y	2	5	9	18	36

Sol

Assume the exponential curve

$$y = a\, b^x$$

Taking common logarithms,

$$Y = \log y = \log a + x \log b = A + Bx$$

Prepare the calculation table:

X	Y	Y = log y	x^2	xY
1	2	0.3010	1	0.3010
2	5	0.6990	4	1.3979
3	9	0.9542	9	2.8627
4	18	1.2553	16	5.0211
5	36	1.5563	25	7.7815
Σx=15	70	ΣY=4.7658	Σ **x²**= 55	ΣxY

From the table,

$$n = 5, \quad \Sigma x = 15, \quad \Sigma Y = 4.7658, \quad \Sigma x^2 = 55, \quad \Sigma xY = 17.3643$$

The normal equations are

$$\Sigma Y = nA + B\Sigma x$$

$$\Sigma xY = A\Sigma x + B\Sigma x^2$$

Substituting the values,

$$4.7658 = 5A + 15B \qquad ...(1)$$

$$17.3643 = 15A + 55B \qquad ...(2)$$

Solving equations (1) and (2), or equivalently using the least squares formula,

$$B = 0.3067$$

$$A = 0.0331$$

Therefore,

$$\log a = A = 0.0331, \quad \log b = B = 0.3067$$

$$a = 10^A = 1.0792, \quad b = 10^B = 2.0262$$

Hence, the fitted exponential curve is

$$y = 1.0792(2.0262)^x$$

3. Fit a curve of the form $y = ab^x$ by the method of least squares for the following data:

X	0	1	2	3	4
Y	3	6	12	24	48

Sol Given curve is $y = ab^x$

Taking logarithms (base 10), $\log y = \log a + x\log b$

Let $Y = \log y, A = \log a, B = \log b$

Then $Y = A + Bx$

x	Y	$Y = \log y$	xY	x^2
0	3	0.4771	0	0
1	6	0.7782	0.7782	1
2	12	1.0792	2.1584	4
3	24	1.3802	4.1406	9
4	48	1.6812	6.7248	16

x	Y	$Y = \log y$	xY	x^2
Σ		5.3959	13.8020	30

Also, $n = 5, \sum x = 10$

Normal equations:

$$\sum Y = nA + B\sum x$$

$$5.3959 = 5A + 10B$$

$$\sum xY = A\sum x + B\sum x^2$$

$$13.8020 = 10A + 30B$$

Solving, $A = 0.4771, B = 0.3010$

Hence $a = \text{antilog}(0.4771) = 3$ and $b = \text{antilog}(0.3010) = 2$

Therefore, $y = 3(2)^x$

Exercise:

1) Fit $y = ab^x$ 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

2) Obtain a relation of the form $y = ab^x$ for the following data by the method of least squares:

X	2	3	4	5	6
Y	8.3	15.4	33.1	65.2	127.4

3) Fit $y = ax^b$ by the method of least squares to the following data:

X	1	2	3	4	5
Y	14	27	40	55	68

4) Obtain a relation of the form $y = ax^b$ for the following data by the method of least squares:

X	2	3	4	5	6
Y	8.3	15.4	33.1	65.2	127.4

5) Fit a power curve (i) $y = ax^b$, (ii) $y = ab^x$

to the following data

x = 1, 2, 3, 4, 5 and y = 2.2, 4.1, 6.7, 9.3, 12.1.

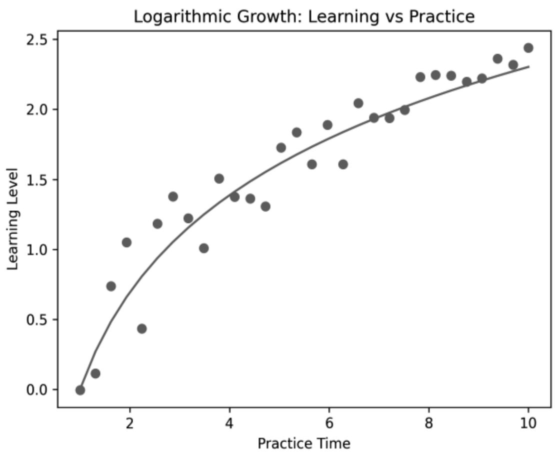
Answers:

1) $A = 1.02115, B = 0.2542892, a = 10.499, b = 1.7959, y = 10.499(1.7959)^x$
2) $A = 0.31, B = 0.30, a = 2.04, b = 1.995, y = 2.04(1.995)^x$
3) $A = 2.6268, B = 0.9852, a = 13.8293, b = 0.9852, y = 13.8293x^{0.9852}$
4) $A = 0.31, B = 0.30, a = 2.04, b = 1.995, y = 2.04x^{1.995}$
5) (i) $y = 2.1049x^{1.0660}$, (ii) $y = 1.6428(1.5263)^x$

3.5. Logarithmic Curve

A logarithmic curve represents a relationship where the increase is rapid at first but slows down over time.

Example: Learning vs practice — learning improves quickly in the beginning and then gradually slows.



Assume the equation: Y = a + b log X

Let X′ = log X

Then Y = a + b X′

Normal Equations

∑y = n a + b ∑X′

∑X′y = a ∑X′ + b ∑(X′²)

Solve to get a and b.

Solved Problems

1. Fit a logarithmic curve of the form $y = a + b\,log\,x$ to the following data:

X	10	30	50	70	90
Y	12	18	21	23	25

Sol

Let $X = log\ x$

Then the logarithmic curve becomes a straight-line equation: $y = a + bX$

X	Y	X = log x	X²	XY
10	12	1.0000	1.0000	12.0000
30	18	1.4771	2.1819	26.5882
50	21	1.6990	2.8865	35.6784
70	23	1.8451	3.4044	42.4373
90	25	1.9542	3.8191	48.8561
Σ	99	7.9754	13.2918	165.5599

Therefore,

$$n = 5,\ \Sigma y = 99.0000,\ \Sigma X = 7.9754,\ \Sigma X^2 = 13.2918,\ \Sigma XY = 165.5599$$

The normal equations are:

$$\Sigma y = na + b\Sigma X$$

$$\Sigma XY = a\Sigma X + b\Sigma X^2$$

Substituting the values,

$$99.0000 = 5a + 7.9754b$$

$$165.5599 = 7.9754a + 13.2918b$$

Solving these simultaneous equations,

$$a = -1.5849, \ b = 13.4067$$

Hence, the required logarithmic curve is,

$$y = -1.5849 + (13.4067) \, log \, x$$

2. Fit a logarithmic curve of the form $y = a + b \, log \, x$ to the following data:

x	5	10	15	20	25
y	2.4	3.1	3.6	4.0	4.3

Sol

Let $X = log \, x$

Then the logarithmic curve becomes a straight-line equation: $y = a + bX$

x	Y	X = log x	X^2	XY
5	2.4	0.6990	0.4886	1.6775
10	3.1	1.0000	1.0000	3.1000
15	3.6	1.1761	1.3832	4.2339
20	4	1.3010	1.6927	5.2041
25	4.3	1.3979	1.9542	6.0111
Σ	17.4000	5.5740	6.5187	20.2267

Therefore,

$$n = 5, \ \Sigma y = 17.4000, \ \Sigma X = 5.5740, \ \Sigma X^2 = 6.5187, \ \Sigma XY = 20.2267$$

The normal equations are:

$$\Sigma y = na + b\Sigma X$$

$$\Sigma XY = a\Sigma X + b\Sigma X^2$$

Substituting the values,

$$17.4000 = 5a + 5.5740b$$

$$20.2267 = 5.5740a + 6.5187b$$

Solving these simultaneous equations,

$$a = 0.4466, \ b = 2.7210$$

Hence, the required logarithmic curve is

$$y = 0.4466 + (2.7210)\, log\, x$$

Exercise:-

1. Fit a logarithmic curve of the form $y = a + b\, log\, x$ to the following data:

x	1	2	4	8	16
y	3.0	4.2	5.1	6.0	6.8

2. Fit a logarithmic curve of the form $\boldsymbol{y} = a + b\, log\, x$ to the following data:

X	10	20	30	40	50
Y	5.2	6.8	7.5	8.1	8.6

Answers:

1) $y = 3.1400 + (3.1226)\, log\, x$

2) $y = 0.4560 + (4.7915)\, log\, x$

Relation between Correlation & Regression

We know that straight line can be denoted as $y = a + bx$ or $x = a + by$. They are known as regression line of y on x or x on y respectively.

These lines of regression can be related to the coefficient of correlation as follows:

$y - \bar{y} = \mathrm{r}\frac{\sigma_y}{\sigma_x}(x - \bar{x})$ ……….(1) (regression line of y on x)

$x - \bar{x} = r\frac{\sigma_x}{\sigma_y}(y - \bar{y}\,)$ ………… (2) (regression line of x on y)

In (1) $\mathrm{r}\frac{\sigma_y}{\sigma_x} = b_{yx}$ known as regression coefficient of y on x

In (2) $r\frac{\sigma_x}{\sigma_y} = b_{xy}$ known as regression coefficient of x on y

$$\mathrm{r}\frac{\sigma_y}{\sigma_x} * r\frac{\sigma_x}{\sigma_y} = b_{yx} * b_{xy}$$

$$r^2 = b_{yx} * b_{xy}$$

$$\text{or} \quad r = \pm\sqrt{(byx * bxy)}$$

- For the above to be true both byx and bxy should have the same sign.
- If both are negative, then we attach a negative sign to r.
- If both are positive, then we attach a positive sign to r.
- The Regression lines passing through $(\bar{x}, \bar{y})$

Solved Problems

1. A company records advertisement expenditure and sales revenue. Find the regression equation of sales on advertisement expenditure and estimate sales when advertisement expenditure is 15 lakhs.

Advertisement X	4	6	8	10	12	14
Sales Y	30	40	50	58	65	72

Sol We know that regression line y on x is

x	y	x^2	xy
4	30	16	120
6	40	36	240
8	50	64	400
10	58	100	580
12	65	144	780
14	72	196	1008
$\sum x = 54$	$\sum y = 315$	$\sum x^2 = 556$	$\sum xy = 3128$

$\bar{x} = 9, \bar{y} = 52.5, \sigma_x = 3.7417,\ \sigma_y = 15.7194$ and r = 0.9963

we use the regression line of y on x is

$$y - \bar{y} = \text{r}\frac{\sigma_y}{\sigma_x}(x - \bar{x})$$

$$y - 52.5 = 0.9963\left(\frac{15.7194}{3.7417}\right)(x - 9)$$

$$y = 4.1856\,x + 14.8296$$

For x = 15 lakhs:

$$y = 4.1856\,(15) + 14.8296$$

$$y = 77.6136$$

2. A damaged statistical record gives the following regression equations and Variance of y is 25.

Regression equations are: $2x - 5y + 5 = 0$ ...(1)

$10x - 9y - 55 = 0$ ...(2)

Find: (i) $\bar{x}$ and $\bar{y}$ (ii) S.D. of x (iii) r.

Sol. Since the two regression lines pass through the point $(\bar{x}, \bar{y})$, Solve the two given equations simultaneously.

$2\bar{x} - 5\bar{y} = -5$ ...(A)

$10\bar{x} - 9\bar{y} = 55$ ...(B)

Solving (A) and (B),

$\bar{x} = 10, \quad \bar{y} = 5$

To identify the two regression coefficients, write each line in the forms y on x and x on y.

The valid pair is selected so that $b_{yx} * b_{xy} = r^2 \leq 1$

From $2x - 5y + 5 = 0$, $y = (2/5)x + 1$, so $b_{yx} = 2/5$

From $10x - 9y - 55 = 0$, $x = (9/10)y + 11/2$, so $b_{xy} = 9/10$

$b_{yx} * b_{xy} = (2/5)(9/10) = 18/50 = 0.36$

$r = +\sqrt{0.36} = 0.6$

$r\frac{\sigma_y}{\sigma_x} = b_{yx} \Rightarrow \sigma_x = 7.5$

Exercise:-

1. The following data show study hours and marks obtained by students. Find the regression equation of marks on study hours and estimate the marks for 9 study hours.

Study Hours X	2	4	5	6	8	10
Marks Y	35	45	50	60	72	85

2. The following regression equations were recovered from an incomplete record.
Variance of x is 64.
Regression equations are:

$$3x + 4y - 72 = 0 \quad ...(1)$$

$$25x + 12y - 344 = 0 \quad ...(2)$$

Find: (i) $\bar{x}$ and $\bar{y}$; (ii) S.D. of y; (iii) r.

Answers:

1. $y = 6.4125\,x + 20.4272$, $y = 78.14$ at $x = 9$

2. $\bar{x} = 8$, $\bar{y} = 12$, $\sigma y = 10$, and $r = -0.6$.

Multiple Linear Regression

Introduction

Multiple Linear Regression is an extension of simple linear regression in which one dependent variable is related to two or more independent variables. It is widely used in economics, engineering, management, agriculture, medicine, and data science for prediction and analysis.

The method helps in understanding how several variables together influence a response variable.

General Form $Y = a + b_1X_1 + b_2X_2 + \cdots + b_nX_n$

where:

- Y= dependent variable
- $X_1, X_2, \ldots, X_n$= independent variables
- a= intercept
- $b_1, b_2, \ldots, b_n$= regression coefficients

Assumptions

1. Linear relationship exists between variables.
2. Errors are independent.
3. Variance of errors is constant.
4. Errors are normally distributed.
5. Independent variables are not perfectly correlated.

Applications

- Sales forecasting
- Weather prediction
- Demand estimation
- Industrial production analysis
- Educational performance analysis

Solved Problems:

1. A company records sales Y based on advertisement expenditure X_1 and number of sales representatives X_2.

X_1	X_2	Y
2	1	20
4	2	32
6	3	44
8	4	56

Fit a multiple linear regression equation.

Sol Using the least squares method, fit $Y = a + b_1X_1 + b_2X_2$

Normal equations are:

$$\sum Y = na + b_1\sum X_1 + b_2\sum X_2$$
$$\sum X_1Y = a\sum X_1 + b_1\sum X_1^2 + b_2\sum X_1X_2$$
$$\sum X_2Y = a\sum X_2 + b_1\sum X_1X_2 + b_2\sum X_2^2$$

X_1	X_2	Y	X_1^2	X_2^2	X_1X_2	X_1Y	X_2Y
2	1	20	4	1	2	40	20
4	2	32	16	4	8	128	64
6	3	44	36	9	18	264	132

X_1	X_2	Y	X_1^2	X_2^2	X_1X_2	X_1Y	X_2Y
8	4	56	64	16	32	448	224
20	10	152	120	30	60	880	440

$$\sum X_1 = 20, \sum X_2 = 10, \sum Y = 152$$
$$\sum X_1^2 = 120, \sum X_2^2 = 30, \sum X_1X_2 = 60$$
$$\sum X_1Y = 880, \sum X_2Y = 440$$

Substituting:

$$152 = 4a + 20b_1 + 10b_2 \qquad (1)$$

$$880 = 20a + 120b_1 + 60b_2 \qquad (2)$$

$$440 = 10a + 60b_1 + 30b_2 \qquad (3)$$

Notice that equation (3) is exactly half of equation (2), because

$$X_1 = 2X_2$$

Hence the variables are perfectly correlated, so unique values of b_1 and b_2 cannot be obtained separately.

Now divide equation (1) by 4: $\quad 38 = a + 5b_1 + \frac{5}{2}b_2 \qquad (4)$

Divide equation (2) by 20: $\quad 44 = a + 6b_1 + 3b_2 \qquad (5)$

Subtract (4) from (5): $\quad 44 - 38 = (a + 6b_1 + 3b_2) - (a + 5b_1 + \frac{5}{2}b_2)$

$$6 = b_1 + \frac{1}{2}b_2$$

Multiply by 2: $\quad 2b_1 + b_2 = 12$

Substitute into (1): $\quad 4a + 10(2b_1 + b_2) = 152$

$$4a + 10(12) = 152$$
$$4a = 32$$
$$a = 8$$

Therefore the regression equation is $Y = 8 + b_1X_1 + b_2X_2$
with $2b_1 + b_2 = 12$

One convenient fitted equation is $Y = 8 + 3X_1 + 6X_2$

which satisfies all observations exactly.

Exercise:

1. Fit a multiple regression equation for the following data:

X_1	X_2	Y
1	2	10
2	3	15
3	5	22
4	6	28

2. The production output Y depends on labour hours X_1 and machine hours X_2.

X_1	X_2	Y
2	1	18
3	2	25
4	3	32
5	4	39

Answers:

1. $Y = 2 + 3X_1 + 2X_2$

2. $Y = 7 + 4X_1 + 3X_2$

Logistic Regression

Logistic Regression is a statistical method used when the dependent variable is categorical in nature. It is commonly used for classification problems where the outcome has two possible categories such as success/failure, pass/fail, or yes/no.

Unlike linear regression, logistic regression predicts probabilities.

The Logistic Regression Model

Suppose Prepresents the probability of success. Then the logistic regression model is

$$P = \frac{1}{1 + e^{-Z}}$$

where

$$Z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n$$

Here,

- P: Probability of occurrence of the event
- $1 - P$: Probability of non-occurrence
- β_0: Intercept term
- $\beta_1, \beta_2, \dots, \beta_n$: Regression coefficients
- $X_1, X_2, \dots, X_n$: Predictor variables

The logistic function produces an S-shaped curve called the sigmoid curve.

Log-Odds or Logit Form

The logistic model can also be written in logarithmic form called the logit model.

$$\ln\left(\frac{P}{1 - P}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n$$

where $\frac{P}{1-P}$ is called the odds ratio.

The term $\ln\left(\frac{P}{1-P}\right)$ is called the log-odds or logit.

Why Do We Use the Logit Form?

Problem with Linear Regression

In ordinary linear regression, $Y = \beta_0 + \beta_1 X$ the predicted values can take any value from $-\infty$ to $+\infty$.

But probabilities must satisfy $0 \leq P \leq 1$
Hence linear regression is not suitable for probability prediction.

Converting Probability into Odds

To overcome this difficulty, probability is transformed into odds.

$$\text{Odds} = \frac{P}{1-P}$$

Example: If $P = 0.8$

Then Odds $= \frac{0.8}{0.2} = 4$
This means the event is four times more likely to occur than not occur.

Taking Logarithm of Odds

Odds values range from 0 to $+\infty$.
To convert this into a quantity ranging from $-\infty$ to $+\infty$
we take the natural logarithm.

$$\ln\left(\frac{P}{1-P}\right)$$

This transformation makes the relationship linear and easier to analyze statistically.

Properties

1. The dependent variable is categorical.
2. The predicted probability always lies between 0 and 1.
3. Logistic regression uses Maximum Likelihood Estimation (MLE) instead of Least Squares Estimation.
4. The logistic curve is S-shaped.
5. It is widely used in classification problems.

Applications

- Medical diagnosis
- Credit risk analysis
- Machine learning classification
- Customer purchase prediction
- Email spam detection

Advantages

1. Suitable for classification.
2. Provides probability estimates.
3. Handles binary outcomes efficiently.
4. Widely used in machine learning.

Limitations

1. Requires large sample size.
2. Sensitive to outliers.
3. Assumes linearity in logit form.

Solved Problem

1. The probability that a student passes an examination depends on study hours X. The fitted logistic model is

$$\log\left(\frac{p}{1-p}\right) = -2 + 0.5X$$

Estimate the probability of passing when $X = 6$ hours.

Sol

Substitute $X = 6$:

$$\log\left(\frac{p}{1-p}\right) = -2 + 0.5(6) = 1$$

Therefore, $\frac{p}{1-p} = e^1 = 2.718$

$$p = \frac{2.718}{1 + 2.718} = 0.731$$

Hence the probability of passing is 0.731 or 73.1%

2. For the model

$$\ln\left(\frac{P}{1-P}\right) = -1 + 0.2X$$

find P when $X = 5$.

Sol

Substituting $X = 5$:

$$\ln\left(\frac{P}{1-P}\right) = -1 + 0.2(5)$$
$$= -1 + 1$$
$$= 0$$

Therefore,

$$\ln\left(\frac{P}{1-P}\right) = 0$$

Taking exponential on both sides,

$$\frac{P}{1-P} = e^0$$
$$\frac{P}{1-P} = 1$$
$$P = 1 - P$$
$$2P = 1$$
$$P = \frac{1}{2} = 0.5$$

Hence, the required probability is 0.5

Exercise

1. A logistic model is given by

$$\log\left(\frac{p}{1-p}\right) = -1 + 0.4X$$

Find the probability when $X = 5$.

2. A logistic model is

$$\log\left(\frac{p}{1-p}\right) = -3 + 0.8X$$

Find the probability when $X = 4$.

3. A company wants to predict whether a customer will buy a product based on monthly income X.

Suppose the fitted logistic regression model is

$$\ln\left(\frac{P}{1-P}\right) = -2 + 0.04X$$

Find the probability when $X = 50$

4. The logistic regression equation is

$$\ln\left(\frac{P}{1-P}\right) = 1 + 0.5X$$

Find the probability when $X = 2$.

Answers:

1. $p = 0.7311$, **2.** $p = 0.5498$, **3.** 0.5, **4.** 0.8808

UNIT 8

Stochastic Processes

Stochastic processes are essential tools for modelling real-world systems where randomness and uncertainty play a major role. From finance to healthcare and AI, they provide a structured way to analyse and predict complex dynamic systems.

Stochastic Process:

A stochastic process is a mathematical model used to describe systems that evolve over time in a random (probabilistic) manner. It studies how things change over time when uncertainty is involved.

Families of random variables which are functions of say time t, are known as Stochastic processes or random processes or random functions. Mathematically, a stochastic process is a set of all random variables {X(t), tϵT} depending on some real parameter like time 't'. The values assumed by random variables X(t) are called 'states and the set of all possible values form the state space of the process.

Example 1: Consider a Random event occurring in time, such as the number of telephone calls received at a switch board. Suppose X(t) is the random variable which represents the number of incoming calls in an interval (0, t) of duration t units. The number of calls within a fixed interval of time, say one unit of time, is a random variable X(1) and the family {X(t), tϵT} constitutes a stochastic process (T = [0, ∞]).

Example 2: Consider a sample experiment like throwing a die. Suppose that X_n is the outcome of the nth throw, n ≥1. Then $\{X_n$, n≥ 1} is a family of random variables such that for a distinct value of n, one gets a distinct random variable X_n ; So $\{X_n$, n≥ 1} constitutes a stochastic process known as Bernoulli process

Example 3: A patient's heart pulse during surgery is measured continuously during interval [0, T]. Stochastic variable X_t represents the occurrence of a heartbeat at time t, 0≤ t ≤ T. Hence, X_t assumes only the values 0 (no heartbeat) and 1 (heartbeat).

A stochastic process is said to be discrete if its state space is discrete. A discrete state process is also called a chain. Also, if the index set t is discrete, then we have a discrete parameter process, otherwise we have a continuous parameter process. A discrete parameter process is also called stochastic sequence $\{X_n\}$, nϵ T.

Types of Stochastic Processes

(a) Based on Time

- **Discrete-time process**: Time steps are separate (e.g., daily stock prices)
- **Continuous-time process**: Time flows continuously (e.g., temperature variation)

(b) Based on State Space

- **Discrete state**: Finite or countable states (e.g., number of customers)
- **Continuous state**: Infinite possible values (e.g., pressure, voltage)

Markov Process (The "Memoryless" Property): If {X(t), t ∈T} is a stochastic process such that, given the value X(s) , the value of X(t), t>s, do not depend on the value of X(u), u<s, then the process is said to be a Markov process.

If for $t_1 < t_2 < t_3 < t_4 < \cdots \ldots \ldots t_n < t$

$$P\{a \leq X(t) \leq b/X(t_1) = (X_1), X(t_2) = (X_2), X(t_3) = (X_3) = \ldots,$$

$$X(t_n) = (X_n)\}$$

$$= P(a \leq X(t) \leq b/X(t_n) = (X_n)\}$$

Then the process $\{X(t), t\epsilon T\}$ is a Markov Process. The future depends **only on the present**, not the past.

Markov Chain: Let X (t) be a Markov process which takes only discrete values whether 't' is discrete or continuous is called the Markov Chain.

Example: Consider the number of heads obtained in the first n tosses of a coin. The possible values are 0,1,2, ...n.

Let S_{n+1} = {x+1 if $(n+1)^{th}$ trial results in head and x, if $(n+1)^{th}$ trial results in tail}

$$S_n = X_0 + X_1 + X_2 + \cdots . +X_n$$

I.e. the conditional probability of S_{n+1} given S_n depends on S_n but not on the manner how the value of S_n is derived.

If we let $P[X_n = a_n/X_{n-1} = a_{n-1}, X_{n-2} = a_{n-2}, X_{n-3} = a_{n-3} \ldots\ldots X_1 = a_1]$ such that

$P[X_n = a_n/X_{n-1} = a_{n-1}]$ for every n

A Markov chain is a stochastic model describing a sequence of possible events in which the probability of each event depends only on the state attained in the previous event. i.e. Consider a system which can be in any one of a finite number of states $E_1, E_2, \ldots \ldots . E_n$. We also assume that the probability of the system being in a given state at the next trial depends only on its present state and not upon the states it may have been in earlier times. If any time, the system is in a state E_i, the probability of it being in the state E_j at the next trial is P_{ij}.

i.e. $P(E_i, E_j) = a_i P_{ij}$

Similarly, $P(E_i, E_j, E_k) = a_i P_{ij} P_{jk}$

Where a_i is the probability distribution of E_i

Here P_{ij} is called the probability of a transition from E_i.

The transition probabilities P_{ij} will be arranged in a matrix form is called matrix of transition probabilities or Transition Probability matrix.

Order of a Markov Chain: A Markov chain is said to be a chain of order one, if for all n,

$$P\{X_n = k / X_{n-1} = j, X_{n-2} = j1, X_{n-3} = j2 \ldots \ldots \ldots ..\}$$

$$= P\{X_n = k / X_{n-1} = j\} = P_{jk}$$

A Markov chain is said to be a chain of order one, if for all n,

$$P\{X_n = k / X_{n-1} = j, X_{n-2} = j1, X_{n-3} = j2 \ldots \ldots \ldots ..\}$$

$$= P\{X_n = k / X_{n-1} = j, X_{n-2} = j1\} = P_{jk}{}^{(2)}$$

Random Walk: A random walk is one of the most important examples of a stochastic process, widely used in mathematics, physics, economics, and computer science.

A **random walk** is a process where a variable moves step-by-step in random directions according to certain probabilities.

Ex. A person standing at 0

At each step:

- Moves **+1 (right)** with probability p
- Moves **−1 (left)** with probability $1 - p$

Mathematical Representation:

$$S_n = X_0 + X_1 + X_2 + \cdots . + X_n$$

Where X_i = +1 with probability P

X_i = -1 with probability 1-P

Types of Random Walk

1. **Simple Symmetric Random Walk**

- p = 0.5
- Equal chance of moving left or right
- No bias

Mean position = 0

2. **Asymmetric Random Walk**

- P ≠ 0.5
- Biased toward one direction

Mean position = n(2p−1)

3. **Random Walk with Absorbing Barriers**

- Stops when it hits a boundary
- Used in **gambler's ruin problem**

4. **Higher-Dimensional Random Walk**

- Movement in 2D or 3D
- Example: particle motion

Stochastic Matrix:

The transition probabilities P_{ij} will be arranged in a matrix of transition probabilities

$$P = \begin{bmatrix} P_{11} & P_{12} & P_{13\ldots} & \cdots\cdots \\ P_{21} & P_{22} & P_{23\ldots} & \cdots\cdots\ldots \\ P_{31} & P_{32} & P_{33\ldots} & \cdots\cdots\ldots \\ \vdots & \vdots & \vdots & \vdots \end{bmatrix}$$

Clearly P is a square matrix with non –negative elements and unit row sums. Such a matrix (finite or infinite) is called a Stochastic matrix

NOTE:

1. A stochastic matrix P is said to be regular if all the entries of some power Pm are positive.

2. A stochastic matrix P is not regular if '1'occurs in the principal main diagonal.

3. If P and Q are stochastic matrices then product PQ is also a stochastic matrix.

Solved Problems:

1. Which of the following matrices are stochastic?

i)$X = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$ ii) $Y = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ iii) $Z = \begin{bmatrix} 0 & 1 \\ 1/2 & 1/4 \end{bmatrix}$ iv) $P = \begin{bmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{bmatrix}$

v) $Q = \begin{bmatrix} 1 & 0 \\ -1 & 1 \end{bmatrix}$

Sol

(i) X is not a square matrix

∴ It is not stochastic

(ii) The matrix is a square matrix with non-negative entries and sum of the elements in each row is equal to 1

∴ Y is stochastic

(iii) Z is a square matrix but the sum in each row is not equal to 1. So, it is not stochastic.

(iv) P is stochastic

(v) Q is not stochastic, because it contains negative elements

2. Which of the following matrices are regular?

$i)\ \ P = \begin{bmatrix} 1/4 & 1/2 & 1/4 \\ 0 & 1 & 0 \\ 1/2 & 0 & 1/2 \end{bmatrix}$ ii) $Q = \begin{bmatrix} 1/2 & 0 & 1/2 \\ 1/2 & 1/2 & 0 \\ 1/4 & 1/4 & 1/2 \end{bmatrix}$ iii) $R = \begin{bmatrix} 0 & 1 & 0 \\ 1/2 & 0 & 1/2 \\ 0 & 0 & 1 \end{bmatrix}$

Sol

i) P is not a regular matrix as 1 lies along the main diagonal.

ii) $Q^2 = Q.Q = \begin{bmatrix} 3/8 & 1/8 & 1/2 \\ 1/2 & 1/4 & 1/4 \\ 3/8 & 1/4 & 3/8 \end{bmatrix}$

Since all the entries of some power of Q are positive, Q is a regular matrix.

iii) $R^2 = R.R = \begin{bmatrix} 1/2 & 0 & 1/2 \\ 0 & 1/2 & 1/2 \\ 0 & 0 & 1 \end{bmatrix}$, $R^3 = R^2.R = \begin{bmatrix} 0 & 1/2 & 1/2 \\ 1/4 & 0 & 3/4 \\ 0 & 0 & 1 \end{bmatrix}$,

We can observe that number of zeros are not at all decreasing, so we can say that R is not a regular matrix.

First order and higher order Markov process:

The Markov chain of the first order is one for which each subsequent state depends only on the immediately preceding one. Markov chains of second or higher orders are the processes in which the next state depends on two or more preceding ones.

n – step transition probabilities:

The state transition probability matrix of a Markov chain gives the probabilities of transitioning from one state to another in a single time unit. Also, define an n - step transition probability matrix $^{(n)}$ whose elements are the n -step transition probabilities.

Chapman–Kolmogorov Equation:

A key finding in Markov process theory is the Chapman-Kolmogorov equation. It connects transition probabilities across several steps.

The likelihood of X_n given X_{n-1} has been examined, by considering unit step transition probabilities. We have taken into consideration the probability of the outcome at the n^{th} step or trial given the outcome at the previous step. p_{ij} provides the probability of a unit step transition from the state I at a trial to the state j at the subsequent trial.

Also k-step transition probability is given by

$$P\{X_{k+n} = j/X_n = i\} = p_{ij}^{(k)} \ldots (1)$$

$p_{ij}^{(k)}$ gives the probability that from the state i at the n^{th} trial, the state j is reached at $(k+n)^{th}$ trial in k steps.

The one step transition probabilities $p_{ij}^{(1)}$ are denoted simply by p_{ij}.

Consider

$$p_{ij}^{(2)} = P\{X_{n+2} = j/X_n = i\} \ldots (2)$$

The state j can be reached from the state i in two steps through some intermediate stage m.

$$\begin{aligned} &P\{X_{n+2} = j, X_{n+1} = m/X_n = i\} \\ &= P\{X_{n+2} = j/X_{n+1} = m, X_n = i\} \cdot P\{X_{n+1} = m/X_n = i\} \\ &= p_{mj}^{(1)} \cdot p_{im}^{(1)} = p_{im}\, p_{mj} \end{aligned}$$

Since these intermediate states $m = 1,2,$ ...are mutually exclusive, we have

$$p_{ij}^{(2)} = P\{X_{n+2} = j/X_n = i\} = \sum_m P\{X_{n+2} = j, X_{n+1} = m/X_n = i\}$$

$$= \sum_m p_{im}\, p_{mj} \ldots (3)$$

By induction, we have

$$p_{ij}^{(k+1)} = P\{X_{n+k+1} = j/X_n = i\}$$

$$= \sum_r P\{X_{n+k+1} = j/X_{n+k} = m\} \cdot P\{X_{n+k} = m/X_n = i\}$$

$$= \sum_m p_{mj}\, p_{im}^{(k)}$$

Similarly, we get

$$p_{ij}^{(k+1)} = \sum_m p_{im}\, p_{mj}^{(k)}$$

In general, we can have

$$p_{ij}^{(k+n)} = \sum_m p_{im}^{(n)}\, p_{mj}^{(k)} = \sum_m p_{im}^{(k)}\, p_{mj}^{(n)} \cdots (4)$$

This equation is a special case of Chapman–Kolmogorov equation, which is satisfied by the transition probabilities of a Markov chain.

For a Markov chain with transition probabilities $Pij^{(n)}$:

$$Pij^{(n+k)} = \sum_m Pim^{(n)} Pmj^{(k)}$$

It means that

Probability of going from state i → j in (n + k) steps = Sum over all possible intermediate states m: (i → m in n steps) × (m→ j in k steps)

It breaks a long transition into smaller parts.

Matrix form: If P is the one-step transition probability matrix, then $P^{(n+k)} = P^{(n)} . P^{(k)}$

It is similar to multiplication rule of probability. It uses Markov property (future depends only on present)

A state vector is a vector (list) that records the probabilities that the system is in any given state at a particular step of the process.

For example, if we know for sure that it is raining today, then the state vector for today will be (1, 0). But tomorrow is another day! We only know there's a 40% chance of rain and 60% chance of no rain, so tomorrow's state vector is (0.4, 0.6).

Classification of states & Chains:

A state is any particular situation that is possible in the system. For example, if we are studying rainy days, then there are two states:

1. It's raining today.

2. It's not raining today.

The system could have many more than two states also.

There are 3 states in Markov process

Accessible State: A state j is said to be accessible from state i if $P_{ij}^{(n)} > 0$ for some $n \geq 1$. This means it is possible to move from state i to state $j$in some number of steps.

Communicating States: States i and j are said to communicate if j is accessible from i, and i is accessible from j. Notation: $i \leftrightarrow j$

Irreducible State: A Markov chain is called irreducible if every state communicates with every other state. That means all states are connected. Otherwise, the chain is called reducible. If the two states i & j are such that each is accessible from the other then it is denoted by

$i \leftrightarrow j$.

Recurrent State: A state i is said to be recurrent if $P_{ii}(n) = 1$ i.e. A state is recurrent if the process returns to it again with probability 1.

Transient State: A state is called transient if, after leaving it, there is a possibility of never returning to it. If there exist j and n such that $P_{ij}(n) > 0$ but $P_{ji}(n) = 0$ for all m, then state 'i' is called transient state.

- The probability that the chain starting from state i returns to state i for the first time at the n^{th} step is denoted by $fii^{(n)}$, n = 1, 2, 3, ... This is called first return time probability (or recurrence time probability).
- If $Fii = \sum fii^{(n)} = 1$, then the return to state i is certain and the state is said to be **persistent or recurrent.**
- If Fii < 1 (return to state i is uncertain), then state i is said to be **transient.**
- The parameter $\mu_i = \Sigma n fii^{(n)}$ is called the **mean recurrence time** of state i.

(a) If μ_i is finite, then state i is said to be **non-null persistent or positive persistent**.

(b) If $\mu_i = \infty$ then state i is said to be **null persistent**.

Absorbing Chain: A state i is called an absorbing state if $P_{ii} = 1$. Once the process enters that state, it never leaves. A Markov chain is absorbing if

i. It has at least one absorbing state.

ii. It is possible to go from every non absorbing state to at least one absorbing state.

Example: $\begin{bmatrix} 1 & 0 & 0 \\ 1/2 & 0 & 1/2 \\ 0 & 0 & 1 \end{bmatrix}$

Here the first and last states are absorbing states.

Periodic and Aperiodic States: The period of state i is $d_i = \gcd\left\{n: P_{ii}^{(n)} > 0\right\}$ where gcd means greatest common divisor.

Periodic State If $d_i > 1$ then the state is periodic.

Aperiodic State If $d_i = 1$ then the state is aperiodic.

Example of Periodicity

Consider the TPM $P = \begin{bmatrix} 0 & 1 & 0 \\ 1/2 & 0 & 1/2 \\ 0 & 1 & 0 \end{bmatrix}$

Possible return steps are: $2,4,6,\ldots$

Therefore, $= \gcd(2,4,6,\ldots) = 2.$

Hence all states are periodic with period 2.

Ergodic Chain: A positive recurrent (positive persistent) and aperiodic state is called **ergodic.** A Markov chain of all such states is called an **ergodic chain**.

i.e. A Markov chain is called **ergodic** if:

1. It is irreducible
2. All states are recurrent
3. All states are aperiodic

So, an ergodic chain is:

- communicative,
- non-periodic,
- stable in the long run.

Short Trick to Classify a Markov Chain

When a problem is given:

Step 1: Check Absorbing States

Look at diagonal entries. If any diagonal entry is 1 → absorbing state exists.

Step 2: Check Communication

See whether every state can reach every other state.

- Yes → irreducible
- No → reducible

Step 3: Find Periodicity

Find all possible return step lengths.

Take gcd.

- gcd = 1 → aperiodic
- gcd > 1 → periodic

Step 4: Final Classification

Property	**Meaning**
Irreducible	All states communicate
Reducible	Some states disconnected
Periodic	gcd > 1
Aperiodic	gcd = 1
Recurrent	Returns surely
Transient	May never return
Ergodic	Irreducible + recurrent + aperiodic

Examples of Irreducible and Ergodic:

Ex.1) State 1 → State 2

State 2 → State 3

State 3 → State 1

This chain is irreducible because you can reach any state from any other state in a finite number of steps.

For example: $1 \to 2$, $2 \to 3$, $3 \to 1$

Here the chain is periodic. If you start in state 1, you can only return to it in 3 steps, 6 steps, 9 steps, etc.

Hence, the period is 3. Since it is periodic, the chain is not ergodic.

Ex. 2) State 1 ↔ State 2 ↔ State 3

This chain is irreducible because you can go from any state to any other state.

It is aperiodic because there is no fixed number of steps required to return to any state.

For example, $P(1 \to 1)$ in one step might be non-zero (self-loop), breaking periodicity.

All the states are positive recurrent, meaning that you will revisit them eventually on average in a finite expected time. Hence, the chain is ergodic.

Solved Problems:

1. Three students X, Y and Z are throwing a ball to each other. X always throws the ball to Y and Y always throws the ball to Z. But Z just as likely to throw the ball to Y as to X. If Z was the first person to throw the ball, find the probability that

(i) X has the ball

(ii) Y has the ball

(iii) Z has the ball after 4 throws.

Sol This is a **Markov chain** problem. Let the states be X, Y, Z.

X → Y (always)

Y → Z (always)

Z → X with probability $1/2$

Z → Y with probability $1/2$

$P11$→ Probability that the ball is thrown from X to X = 0

$P12$→ Probability that the ball is thrown from X to Y = 1

$P13$ = 0, $P21$ = 0, $P22$ = 0, $P23$ = 1, $P31$ = 1/2, $P32$ = 1/2, $P33$ = 0

Step 1: Transition Matrix

$$P = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ \frac{1}{2} & \frac{1}{2} & 0 \end{bmatrix}$$

Step 2: Compute $\boldsymbol{P^2 = P \cdot P}$

$$P^2 = \begin{bmatrix} 0 & 0 & 1 \\ \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & \frac{1}{2} \end{bmatrix}$$

Step 3: Compute $\boldsymbol{P^3 = P^2 \cdot P}$

$$P^3 = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{2} \end{bmatrix}$$

Step 4: Compute $P^4 = P^3 \cdot P$

$$P^4 = \begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{2} \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \end{bmatrix}$$

Step 5: Multiply with Initial Vector

Initial state:

$$\pi^{(0)} = [0 \quad 0 \quad 1]$$

$$\pi^{(4)} = \pi^{(0)}P^4 = \text{(3rd row of } P^4\text{)}$$

$$\pi^{(4)} = \left[\frac{1}{4}, \frac{1}{2}, \frac{1}{4}\right]$$

After 4 throws:

- (i) Probability that X has the ball = 1/4
- (ii) Probability that Y has the ball = 1/2
- (iii) Probability that Z has the ball = ¼

(OR)

This is a **Markov chain** problem.

Let the states be X, Y, Z.

Step 1: Transition Probability Matrix

From the problem:

- $X \to Y = 1$
- $Y \to Z = 1$
- $Z \to X = \frac{1}{2}, Z \to Y = \frac{1}{2}$

So the transition matrix P(in order X, Y, Z) is:

$$P = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ \frac{1}{2} & \frac{1}{2} & 0 \end{bmatrix}$$

Step 2: Initial State Vector

Since **Z starts with the ball**, the initial distribution is:

$$\pi^{(0)} = [0 \quad 0 \quad 1]$$

Step 3: Use Matrix Power Method (P^4)

We compute: $\pi^{(4)} = \pi^{(0)} P^4$

Instead of directly jumping, compute step-by-step:

Step 4: Successive Multiplications

After 1 throw:

$$\pi^{(1)} = [0\ 0\ 1]P = \left[\frac{1}{2}, \frac{1}{2}, 0\right]$$

After 2 throws:

$$\pi^{(2)} = \pi^{(1)} P = \left[0, \frac{1}{2}, \frac{1}{2}\right]$$

After 3 throws:

$$\pi^{(3)} = \left[\frac{1}{4}, \frac{1}{4}, \frac{1}{2}\right]$$

After 4 throws:

$$\pi^{(4)} = \left[\frac{1}{4}, \frac{1}{2}, \frac{1}{4}\right]$$

After 4 throws:

- (i) Probability that X has the ball = 1/4
- (ii) Probability that Y has the ball = 1/2
- (iii) Probability that Z has the ball = 1/4

2. A person owning a scooter has the option to switch over to scooter, bike or a car next time with the probability of [0.3 0.5 0.2] . If the transition probability matrix is

$$P = \begin{bmatrix} 0.4 & 0.3 & 0.3 \\ 0.2 & 0.5 & 0.3 \\ 0.25 & 0.25 & 0.5 \end{bmatrix}$$

what are the probabilities of vehicles related to his fourth purchase.

Sol This is a **Markov chain with 3 states**:

- S = Scooter
- B = Bike
- C = Car

Step 1: Given

Initial distribution: $\pi^{(0)} = [0.3 \;\; 0.5 \;\; 0.2]$

Transition matrix:

$$P = \begin{bmatrix} 0.4 & 0.3 & 0.3 \\ 0.2 & 0.5 & 0.3 \\ 0.25 & 0.25 & 0.5 \end{bmatrix}$$

We need:

$$\pi^{(4)} = \pi^{(0)} P^4$$

Step 2: Compute step by step (accurately)

First purchase: $\pi^{(1)} = \pi^{(0)}P$

$$= [0.3,0.5,0.2]\begin{bmatrix} 0.4 & 0.3 & 0.3 \\ 0.2 & 0.5 & 0.3 \\ 0.25 & 0.25 & 0.5 \end{bmatrix}$$

$$= (0.27,\ 0.39,\ 0.34)$$

Second purchase:

$$\pi^{(2)} = \pi^{(1)}P$$

$$= [0.27,0.39,0.34]\begin{bmatrix} 0.4 & 0.3 & 0.3 \\ 0.2 & 0.5 & 0.3 \\ 0.25 & 0.25 & 0.5 \end{bmatrix}$$

$$= (0.27(0.4) + 0.39(0.2) + 0.34(0.25),\ 0.27(0.3) + 0.39(0.5) + 0.34(0.25),\ 0.27(0.3) + 0.39(0.3) + 0.34(0.5))$$

$$= (0.271,\ 0.361,\ 0.368)$$

Third purchase:

$$\pi^{(3)} = \pi^{(2)}P$$

$$= (0.271,\ 0.361,\ 0.368)\begin{bmatrix} 0.4 & 0.3 & 0.3 \\ 0.2 & 0.5 & 0.3 \\ 0.25 & 0.25 & 0.5 \end{bmatrix}$$

$$= (0.2726,\ 0.3538,\ 0.3736)$$

Fourth purchase:

$$\pi^{(4)} = \pi^{(3)}P$$

$$= [0.26945,0.37615,0.3544]\begin{bmatrix} 0.4 & 0.3 & 0.3 \\ 0.2 & 0.5 & 0.3 \\ 0.25 & 0.25 & 0.5 \end{bmatrix}$$

$$= (0.2726(0.4) + 0.3538(0.2) + 0.3736(0.25),\ 0.2726(0.3) + 0.3538(0.5) + 0.3736(0.25),\ 0.2726(0.3) + 0.3538(0.3) + 0.3736(0.5))$$

$$= (0.2732,\ 0.3521,\ 0.3747)$$

After the 4th purchase:

- Scooter = 0.273
- Bike = 0.352
- Car = 0.375

3. Consider a Markov Process with state space = { 0,1,2} and transition matrix

$$P = \begin{bmatrix} a & b & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} \\ a - \frac{1}{2} & \frac{7}{10} & \frac{1}{5} \end{bmatrix}$$

(i) What can you say about the values of p & q?

(ii) Calculate the transition probability Pij$^{(3)}$

Sol Step 1: Given Transition Matrix

$$P = \begin{bmatrix} a & b & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} \\ a - \frac{1}{2} & \frac{7}{10} & \frac{1}{5} \end{bmatrix}$$

(i) Find values of a and b

Property of Transition Matrix: Each row must sum to **1**

Row 1:

$$a + b + 0 = 1 \Rightarrow a + b = 1 \ldots\ldots\ldots (1)$$

Row 3: From the third row if we sum, all the elements it should be equal to 1

$$\left(a - \frac{1}{2}\right) + \frac{7}{10} + \frac{1}{5} = 1$$

$$a - \frac{1}{2} + \frac{9}{10} = 1$$

So:

$$a + \frac{2}{5} = 1 \Rightarrow a = \frac{3}{5}$$

Now from (1):

$$b = 1 - a = 1 - \frac{3}{5} = \frac{2}{5}$$

Final values: $a = \frac{3}{5}$ $b = \frac{2}{5}$

Step 2: Updated Transition Matrix

$$P = \begin{bmatrix} \frac{3}{5} & \frac{2}{5} & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{10} & \frac{7}{10} & \frac{1}{5} \end{bmatrix}$$

Step 2: Compute $\boldsymbol{P^2 = P \cdot P}$

$$P^2 = \begin{bmatrix} \frac{14}{25} & \frac{6}{25} & \frac{1}{5} \\ \frac{7}{20} & \frac{11}{20} & \frac{1}{10} \\ \frac{21}{50} & \frac{9}{50} & \frac{39}{100} \end{bmatrix}$$

Step 3: Compute $\boldsymbol{P^3 = P^2 \cdot P}$

$$P^3 = \begin{bmatrix} \frac{119}{250} & \frac{91}{250} & \frac{4}{25} \\ \frac{11}{50} & \frac{21}{100} & \frac{59}{200} \\ \frac{381}{1000} & \frac{441}{1000} & \frac{84}{500} \end{bmatrix}$$

$$P^3 \approx \begin{bmatrix} 0.476 & 0.364 & 0.160 \\ 0.220 & 0.210 & 0.295 \\ 0.381 & 0.441 & 0.168 \end{bmatrix}$$

4. The transition probability matrix of a Markov chain is given by

$$P = \begin{bmatrix} 0.4 & 0.6 & 0 \\ 0.2 & 0.3 & 0.5 \\ 0 & 0.3 & 0.7 \end{bmatrix}$$

Is this matrix irreducible?

Sol Given the transition matrix is

$$P = \begin{bmatrix} 0.4 & 0.6 & 0 \\ 0.2 & 0.3 & 0.5 \\ 0 & 0.3 & 0.7 \end{bmatrix}$$

A Markov chain is **irreducible** if:

Every state can be reached from every other state (possibly in multiple steps).

Let states be: 1,2,3

From State 1:

- $1 \to 2$ directly
- $1 \to 3$:
  $1 \to 2 \to 3$

From State 2:

- $2 \to 1$directly
- $2 \to 3$directly

From State 3:

- $3 \to 2$directly
- $3 \to 1$:
  $3 \to 2 \to 1$

Conclusion

Every state communicates with every other state. $Hence$ The Markov chain is irreducible

5. Consider the transition probability matrix $P = \begin{bmatrix} 0 & 0.8 & 0.2 \\ 0.3 & 0.7 & 0 \\ 0.4 & 0.5 & 0.1 \end{bmatrix}$. Classify the Markov chain.

Sol Step 1: Check Absorbing States

A state is absorbing if diagonal entry is 1.

Diagonal entries are: $P_{11} = 0, P_{22} = 0.7, P_{33} = 0.1$

No diagonal entry equals 1.

Hence, there is no absorbing state.

Step 2: Check Communication

- State 1 can reach State 2 and State 3.
- State 2 can reach State 1.
- State 3 can reach State 1 and State 2.

Hence every state communicates with every other state.

Therefore, the chain is irreducible.

Step 3: Check Periodicity

Since self-loops are present: $P_{22} > 0, P_{33} > 0$. Returns are possible in 1 step.

Hence, $d_1 = d_2 = d_3 = 1$. Therefore, all states are aperiodic.

Step 4: Recurrence

All communicating states in a finite irreducible chain are recurrent.

Hence all states are recurrent.

Final Classification

- Irreducible
- Recurrent
- Aperiodic

Therefore, the Markov chain is **Ergodic**.

6. Consider the TPM $P = \begin{bmatrix} 0 & 1 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 1 & 0 \end{bmatrix}$. Classify the chain.

Sol

Step 1: Communication

Each state can reach every other state. Hence the chain is **irreducible**.

Step 2: Find Period

For each state, returns occur only in even number of steps: $2,4,6,\ldots$

Thus, $= \gcd(2,4,6,\ldots) = 2$.

Hence all states are periodic with period 2.

Step 3: Aperiodicity

Since period is not 1, the chain is **not aperiodic**.

Final Classification

- Irreducible
- Recurrent
- Periodic

Therefore, the chain is **not ergodic**.

7. Three boys A, B, C throw a ball among themselves as follows: A always throws to B, B always throws to C and C throws to A or B with equal probability. Construct the TPM and classify the chain.

Sol

From the given data the TPM is

$$P = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ \frac{1}{2} & \frac{1}{2} & 0 \end{bmatrix}$$

Step 1: Communication

- $A \rightarrow B$
- $B \rightarrow C$
- $C \rightarrow A$

Thus every state communicates with every other state.

Hence the chain is **irreducible**.

Step 2: Periodicity

Possible return lengths include: $2,3,4,5, \ldots$

Therefore, $d = \gcd(2,3,4,5, \ldots) = 1$

Hence all states are **aperiodic**.

Step 3: Recurrence

Finite irreducible chains are recurrent.

Thus all states are recurrent.

Final Classification

- Irreducible
- Recurrent
- Aperiodic

Therefore, the chain is **Ergodic and Regular**.

8. A fair die is tossed repeatedly. If $X_n$denotes the maximum number occurring in the first n tosses, find the TPM. Also find $P(X_2 = 6)$.

States: $1,2,3,4,5,6$

$$P = \begin{bmatrix} \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \\ 0 & \frac{2}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \\ 0 & 0 & \frac{3}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \\ 0 & 0 & 0 & \frac{4}{6} & \frac{1}{6} & \frac{1}{6} \\ 0 & 0 & 0 & 0 & \frac{5}{6} & \frac{1}{6} \\ 0 & 0 & 0 & 0 & 0 & \frac{6}{6} \end{bmatrix}$$

Also,

$$P^2 = \frac{1}{36}\begin{bmatrix} 1 & 3 & 5 & 7 & 9 & 11 \\ 0 & 4 & 5 & 7 & 9 & 11 \\ 0 & 0 & 9 & 7 & 9 & 11 \\ 0 & 0 & 0 & 16 & 9 & 11 \\ 0 & 0 & 0 & 0 & 25 & 11 \\ 0 & 0 & 0 & 0 & 0 & 36 \end{bmatrix}$$

$$P(X_2 = 6) = \sum_{i=1}^{6} P(X_2 = 6 \mid X_0 = i)P(X_0 = i)$$

$$= \frac{1}{36}(11+11+11+11+11+36) = \frac{91}{216}$$

9) An urn initially contains 5 black balls and 5 white balls. A ball is drawn from the urn repeatedly.

- If the ball drawn is white, it is put back into the urn.
- Otherwise, it is left out.

Let X_n be the number of black balls remaining after the n^{th} draw.

(i) Show that X_n is a Markov process, (ii) Find the tpm

(iii) Find the transition probabilities (iv) Find P^2 (v) Show that $X_n \to 0$as $n \to \infty$

Sol

(i) Since the number of black balls remaining at stage $n+1$depends only on the number of black balls remaining at stage n, and not on earlier values, $\{X_n\}$is a Markov chain.

(ii) States: 0,1,2,3,4,5

$$P = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{6} & \frac{5}{6} & 0 & 0 & 0 & 0 \\ 0 & \frac{2}{7} & \frac{5}{7} & 0 & 0 & 0 \\ 0 & 0 & \frac{3}{8} & \frac{5}{8} & 0 & 0 \\ 0 & 0 & 0 & \frac{4}{9} & \frac{5}{9} & 0 \\ 0 & 0 & 0 & 0 & \frac{5}{10} & \frac{5}{10} \end{bmatrix}$$

(iii) The non-zero transition probabilities are:

$$P(X_n = 0 \mid X_{n-1} = 0) = 1$$
$$P(X_n = 0 \mid X_{n-1} = 1) = \frac{1}{6}, P(X_n = 1 \mid X_{n-1} = 1) = \frac{5}{6}$$
$$P(X_n = 1 \mid X_{n-1} = 2) = \frac{2}{7}, P(X_n = 2 \mid X_{n-1} = 2) = \frac{5}{7}$$
$$P(X_n = 2 \mid X_{n-1} = 3) = \frac{3}{8}, P(X_n = 3 \mid X_{n-1} = 3) = \frac{5}{8}$$
$$P(X_n = 3 \mid X_{n-1} = 4) = \frac{4}{9}, P(X_n = 4 \mid X_{n-1} = 4) = \frac{5}{9}$$
$$P(X_n = 4 \mid X_{n-1} = 5) = \frac{5}{10}, P(X_n = 5 \mid X_{n-1} = 5) = \frac{5}{10}$$

All remaining probabilities are zero.

(iv) Given

$$P=\begin{bmatrix}1&0&0&0&0&0\\ \frac{1}{6}&\frac{5}{6}&0&0&0&0\\ 0&\frac{2}{7}&\frac{5}{7}&0&0&0\\ 0&0&\frac{3}{8}&\frac{5}{8}&0&0\\ 0&0&0&\frac{4}{9}&\frac{5}{9}&0\\ 0&0&0&0&\frac{5}{10}&\frac{5}{10}\end{bmatrix}$$

$$P^2=P.P=\begin{bmatrix}1&0&0&0&0&0\\ \frac{1}{6}&\frac{5}{6}&0&0&0&0\\ 0&\frac{2}{7}&\frac{5}{7}&0&0&0\\ 0&0&\frac{3}{8}&\frac{5}{8}&0&0\\ 0&0&0&\frac{4}{9}&\frac{5}{9}&0\\ 0&0&0&0&\frac{1}{2}&\frac{1}{2}\end{bmatrix}\begin{bmatrix}1&0&0&0&0&0\\ \frac{1}{6}&\frac{5}{6}&0&0&0&0\\ 0&\frac{2}{7}&\frac{5}{7}&0&0&0\\ 0&0&\frac{3}{8}&\frac{5}{8}&0&0\\ 0&0&0&\frac{4}{9}&\frac{5}{9}&0\\ 0&0&0&0&\frac{1}{2}&\frac{1}{2}\end{bmatrix}$$

$$=\begin{bmatrix}1&0&0&0&0&0\\ \frac{11}{36}&\frac{25}{36}&0&0&0&0\\ \frac{1}{21}&\frac{20}{42}&\frac{25}{49}&0&0&0\\ 0&\frac{3}{28}&\frac{15}{56}+\frac{15}{64}&\frac{25}{64}&0&0\\ 0&0&\frac{1}{6}&\frac{20}{72}+\frac{20}{81}&\frac{25}{81}&0\\ 0&0&0&\frac{2}{9}&\frac{5}{18}+\frac{1}{4}&\frac{1}{4}\end{bmatrix}$$

and then simplifying,

$$P^2 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ \frac{11}{36} & \frac{25}{36} & 0 & 0 & 0 & 0 \\ \frac{1}{21} & \frac{10}{21} & \frac{25}{49} & 0 & 0 & 0 \\ 0 & \frac{3}{28} & \frac{165}{448} & \frac{25}{64} & 0 & 0 \\ 0 & 0 & \frac{1}{6} & \frac{85}{162} & \frac{25}{81} & 0 \\ 0 & 0 & 0 & \frac{2}{9} & \frac{19}{36} & \frac{1}{4} \end{bmatrix}$$

(v) As $n \to \infty$, all black balls are eventually removed.

Hence, $\lim_{n\to\infty} X_n = 0$

and

$$P^\infty = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

10) A gambler has ₹3. He bets ₹1 at a time and wins ₹1 with probability $1/2$. He stops playing if he loses all money or wins enough to reach ₹6. Find:

i. TPM of the Markov chain.

ii. Probability that he has lost all his money at the end of 5 plays.

iii. Probability that the game lasts more than 7 plays.

Sol (i) States 0,1,2,3,4,5,6

where 0 and 6 are absorbing states.

TPM

$$P=\begin{bmatrix}1&0&0&0&0&0&0\\ \frac{1}{2}&0&\frac{1}{2}&0&0&0&0\\ 0&\frac{1}{2}&0&\frac{1}{2}&0&0&0\\ 0&0&\frac{1}{2}&0&\frac{1}{2}&0&0\\ 0&0&0&\frac{1}{2}&0&\frac{1}{2}&0\\ 0&0&0&0&\frac{1}{2}&0&\frac{1}{2}\\ 0&0&0&0&0&0&1\end{bmatrix}$$

(ii) Probability that he loses all money after 5 plays

$P^{(0)}=(0\ \ 0\ \ 0\ \ 1\ \ 0\ \ 0\ \ 0)$

$P^{(1)}=P^{(0)}P=\left(0\ \ 0\ \ \frac{1}{2}\ \ 0\ \ \frac{1}{2}\ \ 0\ \ 0\right)$

$P^{(2)}=P^{(1)}P$

$$=\left(0\ \ 0\ \ \frac{1}{2}\ \ 0\ \ \frac{1}{2}\ \ 0\ \ 0\right)\begin{bmatrix}1&0&0&0&0&0&0\\ \frac{1}{2}&0&\frac{1}{2}&0&0&0&0\\ 0&\frac{1}{2}&0&\frac{1}{2}&0&0&0\\ 0&0&\frac{1}{2}&0&\frac{1}{2}&0&0\\ 0&0&0&\frac{1}{2}&0&\frac{1}{2}&0\\ 0&0&0&0&\frac{1}{2}&0&\frac{1}{2}\\ 0&0&0&0&0&0&1\end{bmatrix}$$

$=\left(0,\frac{1}{4},0,\frac{1}{2},0,\frac{1}{4},0\right)$

similarly $P^{(3)}=(\frac{1}{8},0,\frac{3}{8},0,\frac{3}{8},0,\frac{1}{8})$

$P^{(4)}=(\frac{1}{8},\frac{3}{16},0,\frac{3}{8},0,\frac{3}{16},\frac{1}{8})$

$P^{(5)}=\left(\frac{7}{32},0,\frac{9}{32},0,\frac{9}{32},0,\frac{7}{32}\right)$

Therefore, $P(X_5 = 0) = \frac{7}{32}$

Hence, $P(\text{ruin after 5 plays}) = \frac{7}{32}$

(iii) Probability that the game lasts more than 7 plays

Starting from

$$P^{(5)} = \left(\frac{3}{8}, \frac{5}{32}, 0, \frac{9}{32}, 0, \frac{1}{32}, \frac{1}{16}\right)$$

Calculate $P^{(6)}$

$$P^{(6)} = \left(\frac{7}{32}, \frac{9}{64}, 0, \frac{9}{32}, 0, \frac{9}{64}, \frac{7}{32}\right)$$

$$P^{(7)} = \left(\frac{37}{128}, 0, \frac{27}{128}, 0, \frac{27}{128}, 0, \frac{37}{128}\right)$$

Therefore,

$P(\text{game lasts more than 7 plays})$

$$= P(X_7 = 1) + P(X_7 = 2) + P(X_7 = 3) + P(X_7 = 4) + P(X_7 = 5)$$

$$= 0 + \frac{27}{128} + 0 + \frac{27}{128} + 0$$

$$= \frac{54}{128} = \frac{27}{64}$$

11) Find x, y, z if $\begin{bmatrix} 0 & x & \frac{1}{3} \\ 0 & 0 & y \\ \frac{1}{3} & \frac{1}{4} & z \end{bmatrix}$ is stochastic.

Sol

$$x + \frac{1}{3} = 1$$
$$x = \frac{2}{3}$$
$$y = 1$$
$$\frac{1}{3} + \frac{1}{4} + z = 1$$

$$z = 1 - \frac{7}{12} = \frac{5}{12}$$

So $x = \frac{2}{3}$, $y = 1$, $z = \frac{5}{12}$

12) Find x, y, z if $\begin{bmatrix} 0 & 0.2 & x \\ x & 0.1 & y \\ 0.1 & 0.2 & z \end{bmatrix}$ is stochastic.

Sol

$$\mathrm{x} + 0.2 = 1$$

$$\mathrm{x} = 0.8$$

$$\mathrm{x} + 0.1 + \mathrm{y} = 1$$

$$0.8 + 0.1 + \mathrm{y} = 1$$

$$\mathrm{y} = 0.1$$

$$0.1 + 0.2 + \mathrm{z} = 1$$

$$\mathrm{z} = 0.7$$

So $\mathrm{x} = 0.8$, $\mathrm{y} = 0.1$, $\mathrm{z} = 0.7$

13) An alumni office of a college finds on review that 80% of its students who contribute to annual fund one year will also contribute next year and 30% of those who do not contribute one year will contribute next year. Construct TPM.

Sol

Let State 0 as Contribute and State 1 as Not contribute

$$\mathrm{P}(0 \to 0) = 0.8$$

$$\mathrm{P}(0 \to 1) = 0.2$$

$$\mathrm{P}(1 \to 0) = 0.3$$

$$\mathrm{P}(1 \to 1) = 0.7$$

Hence $\mathrm{P} = \begin{bmatrix} 0.8 & 0.2 \\ 0.3 & 0.7 \end{bmatrix}$

14) Suppose the probability of a dry day (state 0) follows a rainy day (state 1) is $1/3$and the probability of a rainy day follows a dry day is $1/2$. Find the probability that March 3rd and March 5th are dry days given March 1st is dry.

Sol

$$P = \begin{bmatrix} 1/2 & 1/2 \\ 1/3 & 2/3 \end{bmatrix}$$

$$P^2 = \begin{bmatrix} 5/12 & 7/12 \\ 7/18 & 11/18 \end{bmatrix}$$

March 3rd dry day:

$$(P^2)_{00} = \frac{5}{12}$$

$$P^4 = \begin{bmatrix} 413/648 & 235/648 \\ 235/648 & 413/648 \end{bmatrix}$$

March 5th dry day:

$$(P^4)_{00} = \frac{413}{648}$$

15) A raining process is considered as a two-state Markov chain. If it rains, state is 0 and if it does not rain, state is 1. TPM is given by $P = \begin{bmatrix} 0.6 & 0.4 \\ 0.2 & 0.8 \end{bmatrix}$.

Find probability that it will rain for 3 days from today assuming that it is raining today.

Sol

$$P^2 = \begin{bmatrix} 0.44 & 0.56 \\ 0.28 & 0.72 \end{bmatrix}$$

$$P^3 = \begin{bmatrix} 0.376 & 0.624 \\ 0.312 & 0.688 \end{bmatrix}$$

$$0.6 \times 0.44 \times 0.376 = 0.099264$$

So, probability that it will rain for 3 days from today assuming that it is raining today is 0.099264

16) Consider the Markov chain $P = \begin{bmatrix} 3/4 & 1/4 & 0 \\ 1/4 & 1/2 & 1/4 \\ 0 & 3/4 & 1/4 \end{bmatrix}$

Find $P_{01}^{(2)}$ and $P(X_2 = 1, X_0 = 0)$.

Sol

$$P^2 = \begin{bmatrix} 5/8 & 5/16 & 1/16 \\ 5/16 & 1/2 & 3/16 \\ 3/16 & 9/16 & 1/4 \end{bmatrix}$$

$$P_{01}^{(2)} = \frac{5}{16}$$

Given $\qquad P(X_0 = i) = \frac{1}{3}, i = 0,1,2$

$$P(X_2 = 1, X_0 = 0) = P(X_2 = 1 \mid X_0 = 0)\, P(X_0 = 0)$$

$$= \frac{5}{16} \times \frac{1}{3} = \frac{5}{48}$$

17) TPM of a Markov chain having 3 states $P = \begin{bmatrix} 0.1 & 0.5 & 0.4 \\ 0.6 & 0.2 & 0.2 \\ 0.3 & 0.4 & 0.3 \end{bmatrix}$ and $p^{(0)} = (0.7,0.2,0.1)$

(i) Find $P(X_2 = 3)$, (ii) Find $P(X_3 = 2, X_2 = 3, X_1 = 3, X_0 = 2)$

Sol

(i)

$$P^2 = \begin{bmatrix} 0.43 & 0.31 & 0.26 \\ 0.24 & 0.42 & 0.34 \\ 0.36 & 0.35 & 0.29 \end{bmatrix}$$

$$P(X_2 = 3) = \sum_{i=1}^{3} P(X_2 = 3 \mid X_0 = i) P(X_0 = i)$$

$$= (0.26)(0.7) + (0.34)(0.2) + (0.29)(0.1)$$

$$= 0.279$$

(ii) $P(X_3 = 2, X_2 = 3, X_1 = 3, X_0 = 2)$

$P(X_1 = 3, X_0 = 2) = P(X_1 = 3 \mid X_0 = 2)\, P(X_0 = 2)$

$= (0.2)(0.2) = 0.04$

$P(X_2 = 3, X_1 = 3, X_0 = 2) = P(X_2 = 3 \mid X_1 = 3, X_0 = 2)\, P(X_1 = 3, X_0 = 2)$

$= (0.3)(0.04) = 12 \times 10^{-3} = 0.012$

$P(X_3 = 2, X_2 = 3, X_1 = 3, X_0 = 2) = P(X_3 = 2 \mid X_2 = 3, X_1 = 3, X_0 = 2)$

$\times P(X_2 = 3, X_1 = 3, X_0 = 2)$

$= (0.4)(0.012) = 48 \times 10^{-4} = 0.0048$

So $P(X_3 = 2, X_2 = 3, X_1 = 3, X_0 = 2) = 0.0048$

Exercise:

Classify the following Markov chain (1 to 10) with TPM

1. $P = \begin{bmatrix} 0.5 & 0.5 \\ 0.4 & 0.6 \end{bmatrix}$ **2.** $P = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$ **3.** $P = \begin{bmatrix} 1 & 0 & 0 \\ 0.3 & 0.4 & 0.3 \\ 0 & 0.5 & 0.5 \end{bmatrix}$

4. $P = \begin{bmatrix} 0.2 & 0.8 & 0 \\ 0.5 & 0.1 & 0.4 \\ 0.3 & 0.3 & 0.4 \end{bmatrix}$ **5.** $P = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}$ **6.** $P = \begin{bmatrix} 0.6 & 0.4 \\ 0.2 & 0.8 \end{bmatrix}$ **7.** $P = \begin{bmatrix} 0 & 1 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 1 & 0 \end{bmatrix}$

8. $P = \begin{bmatrix} 0.7 & 0.3 & 0 \\ 0.2 & 0.5 & 0.3 \\ 0.1 & 0.4 & 0.5 \end{bmatrix}$ **9.** $P = \begin{bmatrix} 1 & 0 \\ 0.5 & 0.5 \end{bmatrix}$ **10.** $P = \begin{bmatrix} 0 & 0.5 & 0.5 \\ 0.5 & 0 & 0.5 \\ 0.5 & 0.5 & 0 \end{bmatrix}$

11. Consider the Markov chain with TPM $P = \begin{bmatrix} 0.2 & 0.5 & 0.3 \\ 0.4 & 0.3 & 0.3 \\ 0.1 & 0.6 & 0.3 \end{bmatrix}$ and initial probabilities $P(X_0 = 1) = 0.3, P(X_0 = 2) = 0.4, P(X_0 = 3) = 0.3.$

Find $P(X_3 = 1, X_2 = 2, X_1 = 3, X_0 = 2)$.

12. Consider the Markov chain with TPM $P = \begin{bmatrix} 0.3 & 0.4 & 0.3 \\ 0.2 & 0.5 & 0.3 \\ 0.4 & 0.1 & 0.5 \end{bmatrix}$ and initial probabilities $P(X_0 = 1) = 0.2, P(X_0 = 2) = 0.5, P(X_0 = 3) = 0.3.$

Find $P(X_3 = 3, X_2 = 1, X_1 = 2, X_0 = 3)$.

Answers:

1. Irreducible, Aperiodic, Recurrent, Ergodic.

2. Irreducible, Periodic with period 2, Recurrent, Not Ergodic

3. State 1 is absorbing, Reducible, Not Ergodic

4. Irreducible, Aperiodic, Recurrent, Ergodic

5. Irreducible, Periodic with period 3, Recurrent, Not Ergodic

6. Irreducible, Aperiodic, Recurrent, Ergodic

7. Irreducible, Periodic with period 2, Recurrent, Not Ergodic

8. Irreducible, Aperiodic, Recurrent, Ergodic

9. State 1 is absorbing, Reducible, Not Ergodic

10. Irreducible, Aperiodic, Recurrent, Ergodic

11. $P(X_3 = 1, X_2 = 2, X_1 = 3, X_0 = 2) = 0.0288$

12. $P(X_3 = 3, X_2 = 1, X_1 = 2, X_0 = 3) = 0.0018$

Probabilities Of Gambler Ruin:

- A gambler starts with capital x
- Total capital available is k
- At each step:
  - Wins with probability $p \rightarrow$ capital becomes $x + 1$
  - Loses with probability $q = 1 - p \rightarrow$ capital becomes $x - 1$

Define:

$$q_x = \text{Probability of ruin starting from } x$$

Step 1: Recurrence Relation

From one-step analysis:

$$q_x = pq_{x+1} + qq_{x-1}, 1 \leq x \leq k-1$$

Boundary conditions:

$$q_0 = 1, q_k = 0$$

Case 1: Biased Case ($p \neq q$)

General Sol

$$q_x = A + B\left(\frac{q}{p}\right)^x$$

Using boundary conditions:

Final result:

$$\boxed{q_x = \frac{\left(\frac{q}{p}\right)^x - \left(\frac{q}{p}\right)^k}{1 - \left(\frac{q}{p}\right)^k}}$$

Probability of Winning:

$$p_x = 1 - q_x$$

Case 2: Unbiased Case ($p = q = \frac{1}{2}$)

Then equation becomes:

$$2q_x = q_{x+1} + q_{x-1}$$

Sol is linear:

$$q_x = Ax + B$$

Using boundary conditions:

$$\boxed{q_x = \frac{a-x}{a}}$$

Winning Probability:

$$\boxed{p_x = \frac{x}{a}}$$

Expected Duration of the Game:

Case I (Biased Case): $(p \neq q)$

Let d_x denote the expected duration of the game. If the gambler wins in the first trial, then the expected total duration would be $(1{+}d_{x+1})$. If he loses then the expected duration will be $(1{+}d_{x-1})$.

The difference equation will be

$$d_x = p(1 + d_{x+1}) + q(1 + d_{x-1})$$
$$\Rightarrow d_x = 1 + pd_{x+1} + qd_{x-1}\ldots$$

We take the boundary conditions as

$$d_0 = 0, d_k = 0 \text{for } x = 1 \text{ and } x = k - 1\ldots$$

On Solving we get

$$d_x = \frac{k}{q-p} \cdot \frac{1 - \left(\frac{q}{p}\right)^x}{\left(\frac{q}{p}\right)^k - 1} + \frac{x}{q-p}$$

Case II (Unbiased Case): $\boldsymbol{p = q = \frac{1}{2}}$

$$d_x = x(a-x)$$

With initial stakes x and $a-x$, the expected duration is

$$d_x = x(a-x)$$

If the initial stakes are doubled to $2x, 2(a-x)$, then

$$d_{2x} = 2x(2a-2x) = 4x(a-x) = 4d_x$$

Thus, expected duration becomes **4 times** the original when stakes are doubled.

Solved Problems

1. Let $p = \frac{1}{2}, q = \frac{1}{2}, x = 600, k = 1000$

Find expected duration.

Sol $x = x(a-x) = 600(400) = 240000$

2. Calculate probability of ruin and expected duration of the game:

(i) $k = 100, x = 10, p = 0.7$

(ii) $k = 50, x = 30, p = 0.5$

Sol

For (i): Given:

- $k = 100$, $x = 10$, $p = 0.7$, $q = 0.3$
- *Since* $p \neq q$, *use* **biased case formulas**

Step 1: Probability of Ruin

Formula:

$$q_x = \frac{(q/p)^x - (q/p)^k}{1-(q/p)^k}$$

$$\frac{q}{p} = \frac{0.3}{0.7} = 0.428571$$

Now,

$$(0.428571)^{10} \approx 0.000214$$

$$(0.428571)^{100} \approx 0 \text{ (\textit{very small, negligible})}$$

So,

$$q_x \approx \frac{0.000214 - 0}{1-0} = 0.000214$$

Probability of Ruin:

$$\boxed{q_{10} \approx 0.000214}$$

Step 2: Expected Duration

Formula:

$$d_x = \frac{x}{q-p} - \frac{k}{q-p} \cdot \frac{1-(q/p)^x}{1-(q/p)^k}$$

$$q - p = 0.3 - 0.7 = -0.4$$

$$d_x = \frac{10}{-0.4} - \frac{100}{-0.4}(1 - 0.000214)$$

$$= -25 + 250(0.999786)$$

$$= -25 + 249.9465$$

$$= 224.9465$$

Expected Duration:

$$\boxed{d_{10} \approx 224.95 \textit{ steps}}$$

$$p = 0.6, q = 0.4, \frac{q}{p} = \frac{2}{3}$$

For (ii): $p = q = \frac{1}{2}$

$$q_{30} = \frac{k-x}{k} = \frac{50-30}{50} = 0.4$$

$$\mathrm{d}_{30} = \mathrm{x}(\mathrm{a}-\mathrm{x}) = 30(50-30) = 600$$

Expected Duration:

$$\boxed{d_{30} = 600 \textit{ steps}}$$

3. A gambler starts with $x = 3$ units in a game where the total amount is $k = 10$. The probability of winning is $p = 0.4$. Find the probability of ruin and expected duration of the game.

Sol Given $x = 3, k = 10, p = 0.4, q = 0.6$

The probability of ruin when starting with $x$units is

$$q_x = \frac{\left(\frac{q}{p}\right)^x - \left(\frac{q}{p}\right)^k}{1-\left(\frac{q}{p}\right)^k}$$

$$= \frac{(1.5)^3-(1.5)^{10}}{1-(1.5)^{10}} = 0.9581$$

Expected Duration

$$d_x = \frac{x}{q-p} - \frac{k}{q-p} \cdot \frac{1-\left(\frac{q}{p}\right)^x}{1-\left(\frac{q}{p}\right)^k}$$

$$q - p = 0.6 - 0.4 = 0.2$$

Substituting values:

$$d_3 = \frac{3}{0.2} - \frac{10}{0.2}\left(\frac{1 - 3.375}{1 - 57.665}\right)$$

$$= 15 - 50\left(\frac{-2.375}{-56.665}\right)$$

$$= 15 - 50(0.0419)$$

$$= 15 - 2.095$$

$$= 12.905$$

4. A gambler starts with $x = 2$ units and the total amount is $k = 8$. The probability of winning is $p = 0.3$. Find the probability of ruin and expected duration.

Sol Given: $k = 8, x = 2, p = 0.3, q = 0.7$

Since $p \neq q$, use biased case formulas.

Probability of Ruin $$q_x = \frac{\left(\frac{q}{p}\right)^x - \left(\frac{q}{p}\right)^k}{1 - \left(\frac{q}{p}\right)^k}$$

$$\frac{q}{p} = \frac{0.7}{0.3} = \frac{7}{3} \approx 2.3333$$

Now,

$$\left(\frac{7}{3}\right)^2 = \frac{49}{9} \approx 5.444$$

$$\left(\frac{7}{3}\right)^8 \approx 898.15$$

So,

$$q_x = \frac{5.444 - 898.15}{1 - 898.15}$$

$$= \frac{-892.706}{-897.15} \approx 0.995$$

Expected Duration

$$d_x = \frac{x}{q-p} - \frac{k}{q-p} \cdot \frac{1-\left(\frac{q}{p}\right)^x}{1-\left(\frac{q}{p}\right)^k}$$

$$q - p = 0.7 - 0.3 = 0.4$$

Substituting values:

$$d_x = \frac{2}{0.4} - \frac{8}{0.4}\left(\frac{1-5.444}{1-898.15}\right)$$
$$= 5 - 20\left(\frac{-4.444}{-897.15}\right)$$
$$= 5 - 20(0.00495)$$
$$= 5 - 0.099$$
$$= 4.901$$

5. A gambler starts with $x = 4$ units and the total amount is $k = 12$. The probability of winning is $p = 0.5$. Find the probability of ruin and expected duration.

Sol Given $k = 12, x = 4, p = 0.5, q = 0.5$

Since $p = q = \frac{1}{2}$, use fair game formulas.

Probability of Ruin

$$q_x = \frac{k-x}{k}$$

$$q_4 = \frac{12-4}{12} = \frac{8}{12}$$
$$= \frac{2}{3} \approx 0.667$$

Expected Duration

$$d_x = x(k-x)$$

$$d_4 = 4(12-4) = 4(8) = 32$$

6. A gambler starts with $x = 5$ units and the total amount is $k = 15$. The probability of winning is $p = 0.5$. Find the probability of ruin and expected duration.

Sol Given $k = 15, x = 5, p = 0.5, q = 0.5$

Probability of Ruin

$$q_x = \frac{k-x}{k}$$

$$q_5 = \frac{15-5}{15} = \frac{10}{15}$$

$$= \frac{2}{3} \approx 0.667$$

Expected Duration

$$d_x = x(k-x)$$

$$d_5 = 5(15-5) = 5(10) = 50$$

Exercise:

1. A gambler starts with $x = 3$ and total amount $k = 9$. The probability of wining is $p = 0.45$. Find the probability of ruin and expected duration.

2. A gambler starts with $x = 4$ and total amount $k = 11$. The probability of wining is $p = 0.35$. Find the probability of ruin and expected duration.

3. A gambler starts with $x = 2$ and total amount $k = 10$. The probability of wining is

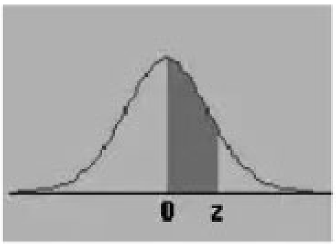
$p = 0.5$. Find the probability of ruin and expected duration.

4. A gambler starts with $x = 6$ and total amount $k = 14$. The probability of wining is $p = 0.5$. Find the probability of ruin and expected duration.

Answers:

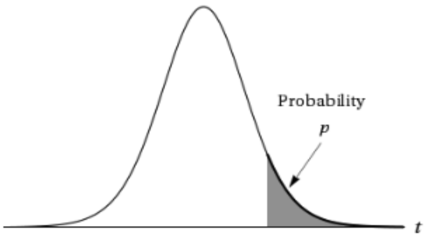
1. $q_z = 0.8376, d_z = 15.39.$ 2. $q_z = 0.9897, d_z = 12.89$

3. $q_z = 0.8, d_z = 16.$ 4. $q_z = \frac{4}{7}, d_z = 48.$



Standard Normal (Z) Table
Area between 0 and z

	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990



t Distribution Critical Values

t Distribution Critical Values

	Upper-tail Probability *p*											
df	**0.25**	**0.20**	**0.15**	**0.10**	**0.05**	**0.025**	**0.02**	**0.01**	**0.005**	**0.0025**	**0.001**	**0.0005**
1	1.000	1.376	1.963	3.078	6.314	12.706	15.895	31.821	63.657	127.321	318.309	636.619
2	0.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.089	22.327	31.599
3	0.765	0.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.215	12.924
4	0.741	0.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.893	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.359	2.764	3.169	3.581	4.144	4.587
11	0.697	0.876	1.088	1.363	1.796	2.201	2.328	2.718	3.106	3.497	4.025	4.437
12	0.695	0.873	1.083	1.356	1.782	2.179	2.303	2.681	3.055	3.428	3.930	4.318
13	0.694	0.870	1.079	1.350	1.771	2.160	2.282	2.650	3.012	3.372	3.852	4.221
14	0.692	0.868	1.076	1.345	1.761	2.145	2.264	2.624	2.977	3.326	3.787	4.140
15	0.691	0.866	1.074	1.341	1.753	2.131	2.249	2.602	2.947	3.286	3.733	4.073
16	0.690	0.865	1.071	1.337	1.746	2.120	2.235	2.583	2.921	3.252	3.686	4.015
17	0.689	0.863	1.069	1.333	1.740	2.110	2.224	2.567	2.898	3.222	3.646	3.965
18	0.688	0.862	1.067	1.330	1.734	2.101	2.214	2.552	2.878	3.197	3.610	3.922
19	0.688	0.861	1.066	1.328	1.729	2.093	2.205	2.539	2.861	3.174	3.579	3.883
20	0.687	0.860	1.064	1.325	1.725	2.086	2.197	2.528	2.845	3.153	3.552	3.850
21	0.686	0.859	1.063	1.323	1.721	2.080	2.189	2.518	2.831	3.135	3.527	3.819
22	0.686	0.858	1.061	1.321	1.717	2.074	2.183	2.508	2.819	3.119	3.505	3.792
23	0.685	0.858	1.060	1.319	1.714	2.069	2.177	2.500	2.807	3.104	3.485	3.768
24	0.685	0.857	1.059	1.318	1.711	2.064	2.172	2.492	2.797	3.091	3.467	3.745
25	0.684	0.856	1.058	1.316	1.708	2.060	2.167	2.485	2.787	3.078	3.450	3.725
26	0.684	0.856	1.058	1.315	1.706	2.056	2.162	2.479	2.779	3.067	3.435	3.707
27	0.684	0.855	1.057	1.314	1.703	2.052	2.158	2.473	2.771	3.057	3.421	3.690
28	0.683	0.855	1.056	1.313	1.701	2.048	2.154	2.467	2.763	3.047	3.408	3.674
29	0.683	0.854	1.055	1.311	1.699	2.045	2.150	2.462	2.756	3.038	3.396	3.659
30	0.683	0.854	1.055	1.310	1.697	2.042	2.147	2.457	2.750	3.030	3.385	3.646
40	0.681	0.851	1.050	1.303	1.684	2.021	2.123	2.423	2.704	2.971	3.307	3.551
50	0.679	0.849	1.047	1.299	1.676	2.009	2.109	2.403	2.678	2.937	3.261	3.496
60	0.679	0.848	1.045	1.296	1.671	2.000	2.099	2.390	2.660	2.915	3.232	3.460
80	0.678	0.846	1.043	1.292	1.664	1.990	2.088	2.374	2.639	2.887	3.195	3.416
100	0.677	0.845	1.042	1.290	1.660	1.984	2.081	2.364	2.626	2.871	3.174	3.390
1000	0.675	0.842	1.037	1.282	1.646	1.962	2.056	2.330	2.581	2.813	3.098	3.300
z^*	0.674	0.842	1.036	1.282	1.645	1.960	2.054	2.326	2.576	2.807	3.090	3.291
	50%	**60%**	**70%**	**80%**	**90%**	**95%**	**96%**	**98%**	**99%**	**99.5%**	**99.8%**	**99.9%**
	Confidence Level *C*											

***F* Critical Values**

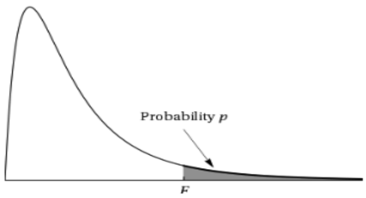


Table -$F_{0.01}$

F - Distribution (α = 0.01 in the Right Tail)

F-table of Critical Values of α = 0.01 for F(df1, df2)

	DF1=1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
DF2=1	4052.18	4999.50	5403.35	5624.58	5763.65	5858.99	5928.36	5981.07	6022.47	6055.85	6106.32	6157.29	6208.73	6234.63	6260.65	6286.78	6313.03	6339.39	6365.86
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	99.45	99.46	99.47	99.47	99.48	99.49	99.50
3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35	27.23	27.05	26.87	26.69	26.60	26.51	26.41	26.32	26.22	26.13
4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55	14.37	14.20	14.02	13.93	13.84	13.75	13.65	13.56	13.46
5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05	9.89	9.72	9.55	9.47	9.38	9.29	9.20	9.11	9.02
6	13.75	10.93	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.72	7.56	7.40	7.31	7.23	7.14	7.06	6.97	6.88
7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.47	6.31	6.16	6.07	5.99	5.91	5.82	5.74	5.65
8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.67	5.52	5.36	5.28	5.20	5.12	5.03	4.95	4.86
9	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.11	4.96	4.81	4.73	4.65	4.57	4.48	4.40	4.31
10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.71	4.56	4.41	4.33	4.25	4.17	4.08	4.00	3.91
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.40	4.25	4.10	4.02	3.94	3.86	3.78	3.69	3.60
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.16	4.01	3.86	3.78	3.70	3.62	3.54	3.45	3.36
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	3.96	3.82	3.67	3.59	3.51	3.43	3.34	3.26	3.17
14	8.86	6.52	5.56	5.04	4.70	4.46	4.28	4.14	4.03	3.94	3.80	3.66	3.51	3.43	3.35	3.27	3.18	3.09	3.00
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.90	3.81	3.67	3.52	3.37	3.29	3.21	3.13	3.05	2.96	2.87
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.55	3.41	3.26	3.18	3.10	3.02	2.93	2.85	2.75
17	8.40	6.11	5.19	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.46	3.31	3.16	3.08	3.00	2.92	2.84	2.75	2.65
18	8.29	6.01	5.09	4.58	4.25	4.02	3.84	3.71	3.60	3.51	3.37	3.23	3.08	3.00	2.92	2.84	2.75	2.66	2.57
19	8.19	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.30	3.15	3.00	2.93	2.84	2.76	2.67	2.58	2.49
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.23	3.09	2.94	2.86	2.78	2.70	2.61	2.52	2.42
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31	3.17	3.03	2.88	2.80	2.72	2.64	2.55	2.46	2.36
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.12	2.98	2.83	2.75	2.67	2.58	2.50	2.40	2.31
23	7.88	5.66	4.77	4.26	3.94	3.71	3.54	3.41	3.30	3.21	3.07	2.93	2.78	2.70	2.62	2.54	2.45	2.35	2.26
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	3.03	2.89	2.74	2.66	2.58	2.49	2.40	2.31	2.21
25	7.77	5.57	4.68	4.18	3.86	3.63	3.46	3.32	3.22	3.13	2.99	2.85	2.70	2.62	2.54	2.45	2.36	2.27	2.17
26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09	2.96	2.82	2.66	2.59	2.50	2.42	2.33	2.23	2.13
27	7.68	5.49	4.60	4.11	3.79	3.56	3.39	3.26	3.15	3.06	2.93	2.78	2.63	2.55	2.47	2.38	2.29	2.20	2.10
28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03	2.90	2.75	2.60	2.52	2.44	2.35	2.26	2.17	2.06
29	7.60	5.42	4.54	4.05	3.73	3.50	3.33	3.20	3.09	3.01	2.87	2.73	2.57	2.50	2.41	2.33	2.23	2.14	2.03
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.84	2.70	2.55	2.47	2.39	2.30	2.21	2.11	2.01
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.67	2.52	2.37	2.29	2.20	2.11	2.02	1.92	1.81
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.50	2.35	2.20	2.12	2.03	1.94	1.84	1.73	1.60
120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47	2.34	2.19	2.04	1.95	1.86	1.76	1.66	1.53	1.38
∞	6.64	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32	2.19	2.04	1.88	1.79	1.70	1.59	1.47	1.33	1.00

***F* Critical Values**

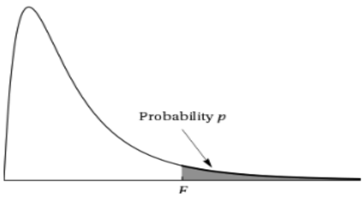
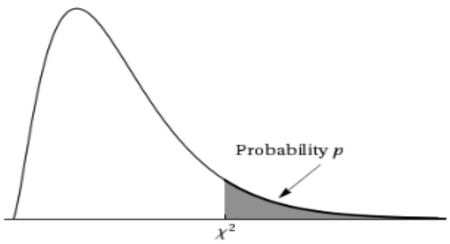


Table -$F_{0.05}$

F-table of Critical Values of α = 0.05 for F(df1, df2)

	DF1=1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
DF2=1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88	243.91	245.95	248.01	249.05	250.10	251.14	252.20	253.25	254.31
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.41	19.43	19.45	19.45	19.46	19.47	19.48	19.49	19.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.37
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.40
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87	1.81
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	1.78
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.81	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79	1.73
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.71
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.15	2.07	1.99	1.95	1.90	1.85	1.80	1.75	1.69
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20	2.13	2.06	1.97	1.93	1.88	1.84	1.79	1.73	1.67
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.12	2.04	1.96	1.91	1.87	1.82	1.77	1.71	1.65
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18	2.10	2.03	1.94	1.90	1.85	1.81	1.75	1.70	1.64
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.58	1.51
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47	1.39
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91	1.83	1.75	1.66	1.61	1.55	1.50	1.43	1.35	1.25
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22	1.00



χ^2 Distribution Critical Values

Right Tail χ^2 Distribution Critical Values

Right Tail Probability p

df	0.25	0.20	0.15	0.10	0.05	0.025	0.02	0.01	0.005	0.0025	0.001	0.0005
1	1.32	1.64	2.07	2.71	3.84	5.02	5.41	6.63	7.88	9.14	10.83	12.12
2	2.77	3.22	3.79	4.61	5.99	7.38	7.82	9.21	10.60	11.98	13.82	15.20
3	4.11	4.64	5.32	6.25	7.81	9.35	9.84	11.34	12.84	14.32	16.27	17.73
4	5.39	5.99	6.74	7.78	9.49	11.14	11.67	13.28	14.86	16.42	18.47	20.00
5	6.63	7.29	8.12	9.24	11.07	12.83	13.39	15.09	16.75	18.39	20.52	22.11
6	7.84	8.56	9.45	10.64	12.59	14.45	15.03	16.81	18.55	20.25	22.46	24.10
7	9.04	9.80	10.75	12.02	14.07	16.01	16.62	18.48	20.28	22.04	24.32	26.02
8	10.22	11.03	12.03	13.36	15.51	17.53	18.17	20.09	21.95	23.77	26.12	27.87
9	11.39	12.24	13.29	14.68	16.92	19.02	19.68	21.67	23.59	25.46	27.88	29.67
10	12.55	13.44	14.53	15.99	18.31	20.48	21.16	23.21	25.19	27.11	29.59	31.42
11	13.70	14.63	15.77	17.28	19.68	21.92	22.62	24.72	26.76	28.73	31.26	33.14
12	14.85	15.81	16.99	18.55	21.03	23.34	24.05	26.22	28.30	30.32	32.91	34.82
13	15.98	16.98	18.20	19.81	22.36	24.74	25.47	27.69	29.82	31.88	34.53	36.48
14	17.12	18.15	19.41	21.06	23.68	26.12	26.87	29.14	31.32	33.43	36.12	38.11
15	18.25	19.31	20.60	22.31	25.00	27.49	28.26	30.58	32.80	34.95	37.70	39.72
16	19.37	20.47	21.79	23.54	26.30	28.85	29.63	32.00	34.27	36.46	39.25	41.31
17	20.49	21.61	22.98	24.77	27.59	30.19	31.00	33.41	35.72	37.95	40.79	42.88
18	21.60	22.76	24.16	25.99	28.87	31.53	32.35	34.81	37.16	39.42	42.31	44.43
19	22.72	23.90	25.33	27.20	30.14	32.85	33.69	36.19	38.58	40.88	43.82	45.97
20	23.83	25.04	26.50	28.41	31.41	34.17	35.02	37.57	40.00	42.34	45.31	47.50
21	24.93	26.17	27.66	29.62	32.67	35.48	36.34	38.93	41.40	43.78	46.80	49.01
22	26.04	27.30	28.82	30.81	33.92	36.78	37.66	40.29	42.80	45.20	48.27	50.51
23	27.14	28.43	29.98	32.01	35.17	38.08	38.97	41.64	44.18	46.62	49.73	52.00
24	28.24	29.55	31.13	33.20	36.42	39.36	40.27	42.98	45.56	48.03	51.18	53.48
25	29.34	30.68	32.28	34.38	37.65	40.65	41.57	44.31	46.93	49.44	52.62	54.95
26	30.43	31.79	33.43	35.56	38.89	41.92	42.86	45.64	48.29	50.83	54.05	56.41
27	31.53	32.91	34.57	36.74	40.11	43.19	44.14	46.96	49.64	52.22	55.48	57.86
28	32.62	34.03	35.71	37.92	41.34	44.46	45.42	48.28	50.99	53.59	56.89	59.30
29	33.71	35.14	36.85	39.09	42.56	45.72	46.69	49.59	52.34	54.97	58.30	60.73
30	34.80	36.25	37.99	40.26	43.77	46.98	47.96	50.89	53.67	56.33	59.70	62.16
40	45.62	47.27	49.24	51.81	55.76	59.34	60.44	63.69	66.77	69.70	73.40	76.09
50	56.33	58.16	60.35	63.17	67.50	71.42	72.61	76.15	79.49	82.66	86.66	89.56
60	66.98	68.97	71.34	74.40	79.08	83.30	84.58	88.38	91.95	95.34	99.61	102.69
80	88.13	90.41	93.11	96.58	101.88	106.63	108.07	112.33	116.32	120.10	124.84	128.26
100	109.14	111.67	114.66	118.50	124.34	129.56	131.14	135.81	140.17	144.29	149.45	153.17