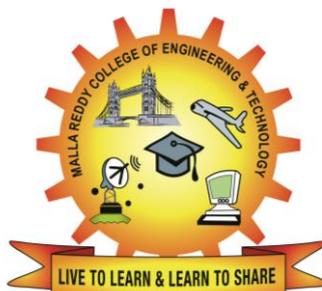# COURSE MATERIAL

II Year B. Tech II- Semester
MECHANICAL ENGINEERING

PROBABILITY AND STATISTICS

R18A0024

MALLA REDDY COLLEGE OF ENGINEERING & TECHNOLOGY

DEPARTMENT OF HUMANITIES & SCIENCES

(Autonomous Institution-UGC, Govt. of India)
Secunderabad-500100,Telangana State, India.
www.mrcet.ac.in

MALLA REDDY COLLEGE OF ENGINEERING & TECHNOLOGY

(Autonomous Institution – UGC, Govt. of India)

DEPARTMENT OF MECHANICAL ENGINEERING

CONTENTS

VISION

- ❖ To establish a pedestal for the integral innovation, team spirit, originality and competence in the students, expose them to face the global challenges and become technology leaders of Indian vision of modern society.

MISSION

- ❖ To become a model institution in the fields of Engineering, Technology and Management.

- ❖ To impart holistic education to the students to render them as industry ready engineers.

- ❖ To ensure synchronization of MRCET ideologies with challenging demands of International Pioneering Organizations.

QUALITY POLICY

- ❖ To implement best practices in Teaching and Learning process for both UG and PG courses meticulously.

- ❖ To provide state of art infrastructure and expertise to impart quality education.

- ❖ To groom the students to become intellectually creative and professionally competitive.

- ❖ To channelize the activities and tune them in heights of commitment and sincerity, the requisites to claim the never - ending ladder of **SUCCESS** year after year.

For more information: www.mrcet.ac.in

VISION

To become an innovative knowledge center in mechanical engineering through state-of-the-art teaching-learning and research practices, promoting creative thinking professionals.

MISSION

The Department of Mechanical Engineering is dedicated for transforming the students into highly competent Mechanical engineers to meet the needs of the industry, in a changing and challenging technical environment, by strongly focusing in the fundamentals of engineering sciences for achieving excellent results in their professional pursuits.

Quality Policy

- ✓ To pursuit global Standards of excellence in all our endeavors namely teaching, research and continuing education and to remain accountable in our core and support functions, through processes of self-evaluation and continuous improvement.

- ✓ To create a midst of excellence for imparting state of art education, industry-oriented training research in the field of technical education.

PROGRAM OUTCOMES

Engineering Graduates will be able to:

1. **Engineering knowledge:** Apply the knowledge of mathematics, science, engineering fundamentals, and an engineering specialization to the solution of complex engineering problems.

2. **Problem analysis**: Identify, formulate, review research literature, and analyze complex engineering problems reaching substantiated conclusions using first principles of mathematics, natural sciences, and engineering sciences.

3. **Design/development of solutions**: Design solutions for complex engineering problems and design system components or processes that meet the specified needs with appropriate consideration for the public health and safety, and the cultural, societal, and environmental considerations.

4. **Conduct investigations of complex problems**: Use research-based knowledge and research methods including design of experiments, analysis and interpretation of data, and synthesis of the information to provide valid conclusions.

5. **Modern tool usage**: Create, select, and apply appropriate techniques, resources, and modern engineering and IT tools including prediction and modeling to complex engineering activities with an understanding of the limitations.

6. **The engineer and society**: Apply reasoning informed by the contextual knowledge to assess societal, health, safety, legal and cultural issues and the consequent responsibilities relevant to the professional engineering practice.

7. **Environment and sustainability**: Understand the impact of the professional engineering solutions in societal and environmental contexts, and demonstrate the knowledge of, and need for sustainable development.

8. **Ethics**: Apply ethical principles and commit to professional ethics and responsibilities and norms of the engineering practice.

9. **Individual and teamwork**: Function effectively as an individual, and as a member or leader in diverse teams, and in multidisciplinary settings.

10. **Communication**: Communicate effectively on complex engineering activities with the engineering community and with society at large, such as, being able to comprehend and write effective reports and design documentation, make effective presentations, and give and receive clear instructions.

11. **Project management and finance**: Demonstrate knowledge and understanding of the engineering and management principles and apply these to one's own work, as a member and leader in a team, to manage projects and in multidisciplinary environments.

12. **Life-long learning**: Recognize the need for and have the preparation and ability to engage in independent and life-long learning in the broadest context of technological change.

PROGRAM SPECIFIC OUTCOMES (PSOs)

PSO1   Ability to analyze, design and develop Mechanical systems to solve the Engineering problems by integrating thermal, design and manufacturing Domains.

PSO2   Ability to succeed in competitive examinations or to pursue higher studies or research.

PSO3   Ability to apply the learned Mechanical Engineering knowledge for the Development of society and self.

Program Educational Objectives (PEOs)

The Program Educational Objectives of the program offered by the department are broadly listed below:

PEO1: PREPARATION

To provide sound foundation in mathematical, scientific and engineering fundamentals necessary to analyze, formulate and solve engineering problems.

PEO2: CORE COMPETANCE

To provide thorough knowledge in Mechanical Engineering subjects including theoretical knowledge and practical training for preparing physical models pertaining to Thermodynamics, Hydraulics, Heat and Mass Transfer, Dynamics of Machinery, Jet Propulsion, Automobile Engineering, Element Analysis, Production Technology, Mechatronics etc.

PEO3: INVENTION, INNOVATION AND CREATIVITY

To make the students to design, experiment, analyze, interpret in the core field with the help of other inter disciplinary concepts wherever applicable.

PEO4: CAREER DEVELOPMENT

To inculcate the habit of lifelong learning for career development through successful completion of advanced degrees, professional development courses, industrial training etc.

PEO5: PROFESSIONALISM

To impart technical knowledge, ethical values for professional development of the student to solve complex problems and to work in multi-disciplinary ambience, whose solutions lead to significant societal benefits.

Blooms Taxonomy

Bloom's Taxonomy is a classification of the different objectives and skills that educators set for their students (learning objectives). The terminology has been updated to include the following six levels of learning. These 6 levels can be used to structure the learning objectives, lessons, and assessments of a course.

1. **Remembering**: Retrieving, recognizing, and recalling relevant knowledge from long- term memory.
2. **Understanding**: Constructing meaning from oral, written, and graphic messages through interpreting, exemplifying, classifying, summarizing, inferring, comparing, and explaining.
3. **Applying**: Carrying out or using a procedure for executing or implementing.
4. **Analyzing**: Breaking material into constituent parts, determining how the parts relate to one another and to an overall structure or purpose through differentiating, organizing, and attributing.
5. **Evaluating**: Making judgments based on criteria and standard through checking and critiquing.
6. **Creating**: Putting elements together to form a coherent or functional whole; reorganizing elements into a new pattern or structure through generating, planning, or producing.
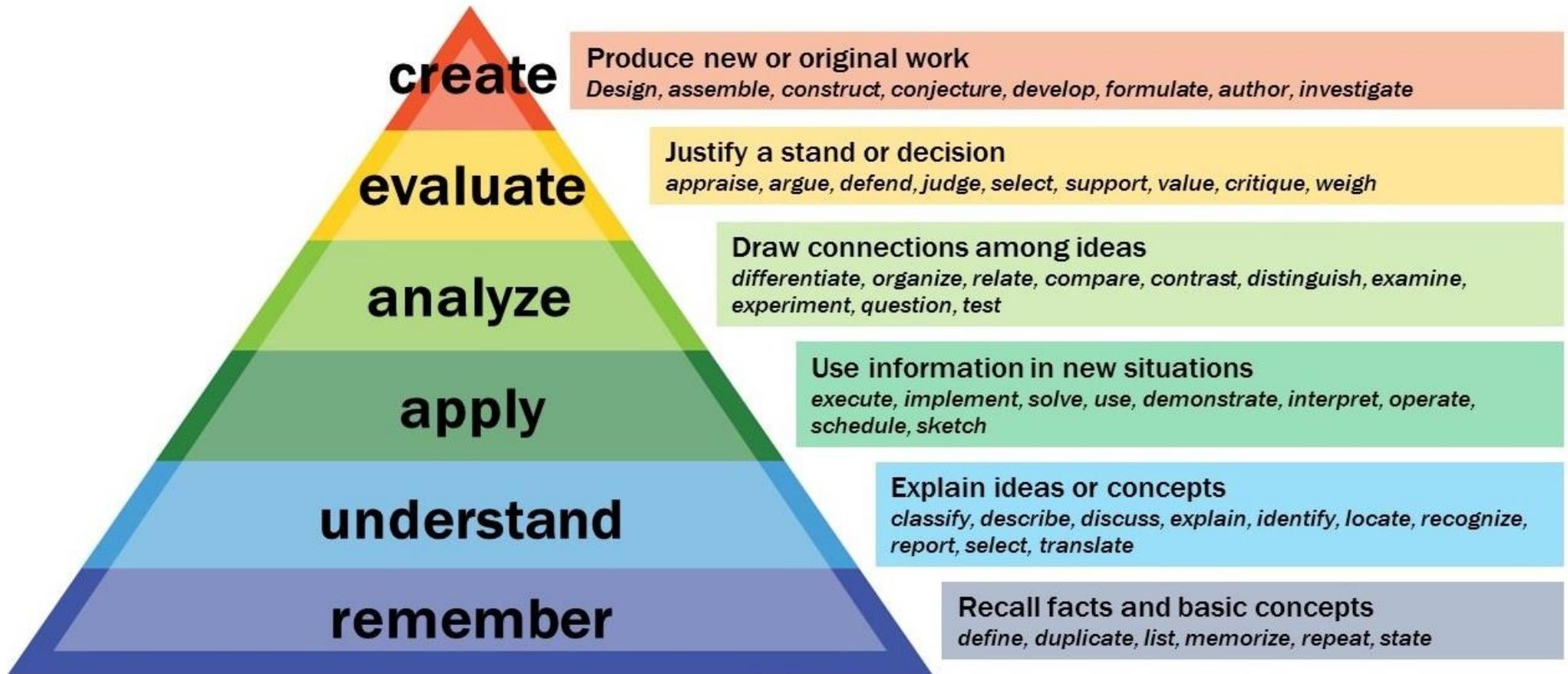
(R18A0024) PROBABILTY AND STATISTICS
(Common to CSE, IT, ANE and MECH)

UNIT – I: Random Variables

Single and multiple Random variables -Discrete and Continuous. Probability distribution function, mass function and density function of probability distribution. mathematical expectation and variance.

UNIT-II: Probability distributions

 Binomial distribution – properties, mean and variance, Poisson distribution – properties, mean and variance and Normal distribution – properties, mean and variance.

UNIT -III :Correlation and Regression

Correlation -Coefficient of correlation, Rank correlation, Regression- Regression Coefficients , Lines of Regression.

UNIT –IV: Sampling

Sampling: Definitions of population, sampling, statistic, parameter - Types of sampling - Expected values of sample mean and variance, Standard error - Sampling distribution of means and variance. Estimation - Point estimation and Interval estimation.

Testing of hypothesis: Null and Alternative hypothesis - Type I and Type II errors, Critical region - confidence interval - Level of significance, One tailed and Two tailed test.

Unit-V: Statistical Inference

Large sample Tests: Test of significance - Large sample test for single mean, difference of means, single proportion, difference of proportions.
Small samples: Test for single mean, difference of means, test for ratio of variances (F-test) - Chi-square test for goodness of fit and independence of attributes.

Suggested Text Books:

i)   Fundamental of Statistics by S.C. Gupta ,7th Edition,2016.
ii)   Fundamentals of Mathematical Statistics by SC Gupta and V.K. Kapoor
iii)   B.S. Grewal, Higher Engineering Mathematics, Khanna Publishers, 35th Edition, 2000.

References :

i)   Probability and Statistics for Engineers and Scientists by Sheldon M.Ross,Academic Press.
ii)   Probability and Statistics by T.K..V Iyengar& B.Krishna Gandhi S.Ranganatham ,MVSSAN Prasad. S Chand Publishers

Course Outcomes:

- Understand a random variable that describes randomness or an uncertainty in certain realistic situation. It can be either discrete or continuous type.

- In the discrete case, study of the binomial and the Poisson random variables and the normal random variable for the continuous case predominantly describe important probability distributions. Important statistical properties for these random variables provide very good insight and are essential for industrial applications.

- Make data-driven decisions by using correlation and regression.

- The types of sampling, Sampling distribution of means, Sampling distribution of variance, Estimations of statistical parameters, Testing of hypothesis of few unknown statistical parameters.

- Draw statistical inference using samples of a given size which is taken from a population and to apply statistical methods for analyzing experimental data.

.
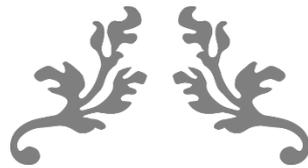
UNIT 1

RANDOM VARIABLES

OBJECTIVES

1. To understand a random variable that describes randomness or an uncertainty in certain realistic situations which can be either discrete or continuous type.

2. To learn functions of multiple random variables through joint distributions since the random situations are described as functions of multiple random variables.

OUTCOME:

Describe randomness in certain realistic situation which can be either discrete or continuous type.

UNIT – I

RANDOM VARIABLES

Random Variable

A Random Variable X is a real valued function from sample space S to a real number R.

(or)

A Random Variable X is a real number which is determined by the outcomes of the random experiment.

Eg:1.Tosssing 2 coins simultaneously

   Sample space ={HH,HT,TH,TT}

Let the random variable be getting number of heads then

    X(S)={0,1,2}.

2.Sum of the two numbers on throwing 2 dice

    X(S)={2,3,4,5,6,7,8,9,10,11,12}.

Types of Random Variables:

1.**Discrete Random Variables :** A Random Variable X is said to be discrete if it takes only the values of the set {0,1,2…..n}.

Eg:1.Tosssing a coin, throwing a dice,number of defective items in a bag.

2.Continuous Random Variables: A Random Variable X which takes all possible values in a given interval of domain.

Eg: Heights, weights of students in a class.

Discrete Probability Distribution:

Let x is a Discrete Random Variable with possible outcomes $x_1, x_2, x_3 \ldots x_n$ having probabilities $p(x_i) for\ i = 1,2 \ldots n$ .If $p(x_i) > 0\ and\ \sum_{i=1}^{n} p(x_i) = 1$ then the function $p(x_i)$ is called **Probability mass function** of a random variable X and $\{ x_i, p(x_i)\}\ for\ i = 1,2 \ldots n$ is called **Discrete Probability Distribution.**

Eg: Tossing 2 coins simultaneously

   Sample space ={HH,HT,TH,TT}

Let the random variable be getting number of heads then

    X(S)={0,1,2}.

Probability of getting no heads $= \frac{1}{4}$, Probability of getting 1 head $= \frac{1}{2}$, Probability of getting 2 heads $= \frac{1}{4}$

∴ Discrete Probability Distribution is

x_i	0	1	2
$p(x_i)$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

Cumulative Distribution function is given by $F(x) = p[X \leq x] = \sum_{i=0}^{x} p(x_i)$.

Properties of Cumulative Distribution function:

1. $P[a < x < b] = F(b) - F(a) - P[X = b]$

2. $P[a \leq x \leq b] = F(b) - F(a) - P[X = a]$

3. $P[a < x \leq b] = F(b) - F(a)$

4. $P[a \leq x < b] = F(b) - F(a) - P[X = b] + P[X = a]$

Mean: The mean of the discrete Probability Distribution is defined as

$$\mu = \frac{\sum_{i=1}^{n} x_i p(x_i)}{\sum_{i=1}^{n} p(x_i)} = \sum_{i=1}^{n} x_i p(x_i) \text{ since } \sum_{i=1}^{n} p(x_i) = 1$$

Expectation: The Expectation of the discrete Probability Distribution is defined as

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

In general, $E(g(x)) = \sum_{i=1}^{n} g(x_i) p(x_i)$

Properties:

1) $E(X) = \mu$

2) $E(X) = k\, E(X)$

3) $E(X + k) = E(X) + k$

4)) $E(aX \pm b) = aE(X) \pm b$

Variance: The variance of the discrete Probability Distribution is defined as

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

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

$$= \sum x_i^2 p_i - \mu^2$$

Properties:

1) $V(c) = 0$ where c is a constant

2) $V(kX) = k^2\, V(X)$

3) $V(X + k) = V(X)$

4) $V(aX \pm b) = a^2\, V(X)$

Problems

1.If 3 cars are selected randomly from 6 cars having 2 defective cars.

a)Find the Probability distribution of defective cars.

b)Find the Expected number of defective cars.

Sol: Number of ways to select 3 cars from 6 cars $= 6_{c_3}$

Let random variable X(S) = Number of defective cars = {0,1,2}

Probability of non defective cars $= \dfrac{4_{c_3}\,2_{c_0}}{6_{c_3}} = \dfrac{1}{5}$

Probability of one defective cars $= \dfrac{4_{c_2}\,2_{c_1}}{6_{c_3}} = \dfrac{3}{5}$

Probability of two defective cars $= \dfrac{4_{c_1}\,2_{c_2}}{6_{c_3}} = \dfrac{1}{5}$

Clearly , $p(x_i) > 0\ and\ \sum_{i=1}^{n} p(x_i) = 1$

Probability distribution of defective cars is

x_i	0	1	2
$p(x_i)$	$\dfrac{1}{5}$	$\dfrac{3}{5}$	$\dfrac{1}{5}$

Expected number of defective cars $= \sum_{i=1}^{n} x_i p(x_i) = 0\left(\dfrac{1}{5}\right) + 1\left(\dfrac{3}{5}\right) + 2\left(\dfrac{1}{5}\right) = 1$

2.Let X be a random variable of sum of two numbers in throwing two fair dice. Find the probability distribution of X, mean ,variance.

Sol: Sample space of throwing two dices is

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

(2,1),(2,2),(2,3),(2,4),(2,5),(2,6)

(3,1),(3,2),(3,3),(3,4),(3,5),(3,6)

(4,1),(4,2),(4,3),(4,4),(4,5),(4,6)

(5,1),(5,2),(5,3),(5,4),(5,5),(5,6)

(6,1),(6,2),(6,3),(6,4),(6,5),(6,6)}

$\therefore n(S) = 36.$

Let X = Sum of two numbers in throwing two dice = {2,3,4,5,6,7,8,9,10,11,12}

X	Favorable cases	No of Favorable cases	$p(x)$
2	(1,1)	1	$\dfrac{1}{36}$
3	(2,1),)(1,2)	2	$\dfrac{2}{36}$
4	(3,1),(2,2),(1,3)	3	$\dfrac{3}{36}$
5	(4,1),(3,2),(2,3),(1,4)	4	$\dfrac{4}{36}$
6	(5,1),(4,2),(3,3),(2,4),(1,5)	5	$\dfrac{5}{36}$
7	(6,1),(5,2),(4,3),(3,4),(2,5),(1,6)	6	$\dfrac{6}{36}$
8	(6,2),(5,3),(4,4),(3,5),(2,6)	5	$\dfrac{5}{36}$
9	(6,3),(5,4),(4,5),(3,6)	4	$\dfrac{4}{36}$
10	(6,4),(5,5),(4,6)	3	$\dfrac{3}{36}$
11	(6,5),(5,6)	2	$\dfrac{2}{36}$
12	(6,6)	1	$\dfrac{1}{36}$

Clearly , $p(x_i) > 0$ and $\sum_{i=1}^{n} p(x_i) = 1$

Probability distribution   is given by

x_i	2	3	4	5	6	7	8	9	10	11	12
$p(x_i)$	$\dfrac{1}{36}$	$\dfrac{2}{36}$	$\dfrac{3}{36}$	$\dfrac{4}{36}$	$\dfrac{5}{36}$	$\dfrac{6}{36}$	$\dfrac{5}{36}$	$\dfrac{4}{36}$	$\dfrac{3}{36}$	$\dfrac{2}{36}$	$\dfrac{1}{36}$

$$\text{Mean} = \mu = \sum_{i=1}^{n} x_i p(x_i) = 2\left(\frac{1}{36}\right) + 3\left(\frac{2}{36}\right) + 4\left(\frac{3}{36}\right) + 5\left(\frac{4}{36}\right) + 6\left(\frac{5}{36}\right) + 7\left(\frac{6}{36}\right) +$$

$$8\left(\frac{5}{36}\right) + 9\left(\frac{4}{36}\right) + 10\left(\frac{3}{36}\right) + 11\left(\frac{2}{36}\right) + 12\left(\frac{1}{36}\right) = 7.$$

Variance $= V(X) = \sum x_i^2 p_i - \mu^2$

$$= 4\left(\frac{1}{36}\right) + 9\left(\frac{2}{36}\right) + 16\left(\frac{3}{36}\right) + 25\left(\frac{4}{36}\right) + 36\left(\frac{5}{36}\right) + 49\left(\frac{6}{36}\right) + 64\left(\frac{5}{36}\right) +$$
$$81\left(\frac{4}{36}\right) + 100\left(\frac{3}{36}\right) + 121\left(\frac{2}{36}\right) + 144\left(\frac{1}{36}\right) - 49 = 5.83$$

3. Let X be a random variable of maximum of two numbers in throwing two fair dice simultaneously. Find the

a)probability distribution of X

b)mean

c)variance

d)P(1<x<4)

e)P(2≤ x ≤ 4).

Sol: Sample space of throwing two dices $= S =\{(1,1),(1,2),(1,3),(1,4),(1,5),(1,6)$

$(2,1),(2,2),(2,3),(2,4),(2,5),(2,6)$

$(3,1),(3,2),(3,3),(3,4),(3,5),(3,6)$

$(4,1),(4,2),(4,3),(4,4),(4,5),(4,6)$

$(5,1),(5,2),(5,3),(5,4),(5,5),(5,6)$

$(6,1),(6,2),(6,3),(6,4),(6,5),(6,6)\}$

∴ $n(S) = 36$.

Let X = Maximum of two numbers in throwing two dice $= \{1,2,3,4,5,6,\}$

X	Favorable cases	No of Favorable cases	$p(x)$
1	(1,1)	1	$\frac{1}{36}$
2	(2,1),)(1,2),(2,2)	3	$\frac{3}{36}$
3	(3,1),(1,3),(2,3)(3,3),(3,2)	5	$\frac{5}{36}$
4	(1,4),(4,1),(4,2),(2,4)(4,3),(3,4),(4,4)	7	$\frac{7}{36}$
5	(1,5),(5,1),(2,5),(5,2)(3,5),(5,3),(5,4),(4,5),(5,5)	9	$\frac{9}{36}$
6	(1,6)(6,1),(6,2),(2,6),(6,3),(3,6),(4,6),(6,4),(6,5)(5,6),(6,6)	11	$\frac{11}{36}$

Clearly , $p(x_i) > 0$ and $\sum_{i=1}^{n} p(x_i) = 1$

Probability distribution   is given by

x_i	1	2	3	4	5	6
$p(x_i)$	$\dfrac{1}{36}$	$\dfrac{3}{36}$	$\dfrac{5}{36}$	$\dfrac{7}{36}$	$\dfrac{9}{36}$	$\dfrac{11}{36}$

$$\text{Mean} = \mu = \sum_{i=1}^{n} x_i p(x_i) = 1\left(\frac{1}{36}\right) + 2\left(\frac{3}{36}\right) + 3\left(\frac{5}{36}\right) + 4\left(\frac{7}{36}\right) + 5\left(\frac{9}{36}\right) + 6\left(\frac{11}{36}\right)$$

$$= 4.4\,7.$$

Variance $= V(X) = \sum x_i^2 p_i - \mu^2$

$$= 1\left(\frac{1}{36}\right) + 4\left(\frac{3}{36}\right) + 9\left(\frac{5}{36}\right) + 16\left(\frac{7}{36}\right) + 25\left(\frac{9}{36}\right) + 36\left(\frac{11}{36}\right)$$

$$\therefore \text{Variance} = 1.99.$$

4.A random variable X has the following probability function

x_i	-3	-2	-1	0	1	2	3
$p(x_i)$	k	0.1	k	0.2	2k	0.4	2k

Find k ,mean, variance.

Sol: We know that $\sum_{i=1}^{n} p(x_i) = 1$

i.e k+0.1+k+0.2+2k+0.4+2k = 1

i.e 6k+0.7 = 1 $\therefore k = 0.05$

$$\text{Mean} = \mu = \sum_{i=1}^{n} x_i p(x_i) = k(-3) + 0.1(-2) + k(-1) + 2k(1) + 2(0.4) + 3(2k)$$

$$= 0.8.$$

Variance $= V(X) = \sum x_i^2 p_i - \mu^2$
$$= k(-3)^2 + 0.1^2(-2) + k(-1)^2 + 2k(1) + 4(0.4) + 9(2k)$$

$$\therefore \text{Variance} = 2.86.$$

Continuous Probability distribution:

Let X be a continuous random variable taking values on the interval (a,b). A function f(x) is said to be the Probability density function of x if

  i) $f(x) > 0 \;\forall\; x \in (a, b)$

  ii)   Total area under the probability curve is 1 i. e, $\int_{a}^{b} f(x)dx = 1.$

  iii)   For two distinct numbers 'c' and 'd' in (a, b) is given by $P(c < x < d) = $ Area under the probability curve between ordinates x = c and x = d i. e $\int_{c}^{d} f(x)dx.$

Note: $P(c < x < d) = P(c \le x \le d) = P(c \le x < d) = P(c < x \le d)$

Cumulative distribution function of $f(x)$ is given by

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

Mean: The mean of the continuous Probability Distribution is defined as

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

Expectation: The Expectation of the continuous Probability Distribution is defined as

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

In general, $E(g(x)) = \int_{-\infty}^{\infty} g(x) \, f(x)dx.$

Properties:

1) $E(X) = \mu$

2) $E(X) = k \, E(X)$

3) $E(X + k) = E(X) + k$

4)) $E(aX \pm b) = aE(X) \pm b$

Variance: The variance of the Continuous Probability Distribution is defined as

$Var(X) = V(X) = \int_{-\infty}^{\infty} x^2 \, f(x)dx - \mu^2.$

Properties:

1) $V(c) = 0$ where c is a constant

2) $V(kX) = k^2 \, V(X)$

3) $V(X + k) = V(X)$

4) $V(aX \pm b) = a^2 \, V(X)$

Mean Deviation: Mean deviation of continuous probability distribution function is defined as

$$\int_{-\infty}^{\infty} |x - \mu| \, f(x)dx.$$

Median: Median is the point which divides the entire distribution in to two equal parts. In case of continuous distribution ,median is the point which divides the total area in to two equal parts i.e, $\int_{a}^{M} f(x)dx = \int_{M}^{b} f(x)dx = \frac{1}{2} \; \forall \, x \in (a, b)$.

Mode: Mode is the value of x for which f(x) is maximum.

i.e $f'(x) = 0$ and $f''(x) < 0$ for $x \in (a, b)$

Problems

1. If the probability density function $f(x) = \dfrac{k}{1+x^2}$ $\quad -\infty < x < \infty$. Find the value of 'k' and probability distribution function of $f(x)$.

Sol: Since total area under the probability curve is 1 i. e, $\int_a^b f(x)dx = 1$.

$$\int_{-\infty}^{\infty} \frac{k}{1+x^2} \, dx = 1.$$
$$2k(\tan^{-1} x)\Big|_0^{\infty} = 1$$
$$2k(\tan^{-1}\infty - \tan^{-1}0) = 1$$
$$\therefore k = \frac{1}{\pi}$$

Cumulative distribution function of f(x) is given by

$$\int_{-\infty}^{x} f(x)dx = \int_{-\infty}^{x} \frac{k}{1+x^2}\,dx = \frac{1}{\pi}(\tan^{-1} x)\Big|_{-\infty}^{x} = \frac{1}{\pi}\left[\frac{\pi}{2} + (\tan^{-1} x)\right].$$

2. If the probability density function $f(x) = ce^{-|x|}$ $\quad -\infty < x < \infty$.

Find the value of 'c', mean and variance.

Sol: Since total area under the probability curve is 1 i. e, $\int_a^b f(x)dx = 1$.

$$\int_{-\infty}^{\infty} ce^{-|x|}\,dx = 1$$
$$2\int_{0}^{\infty} ce^{-x}\,dx = 1$$
$$2c\left(\frac{e^{-x}}{-1}\right)\Big|_0^{\infty} = 1$$
$$\therefore c = \frac{1}{2}$$

Mean$=\mu = \int_{-\infty}^{\infty} x\, f(x)dx = \frac{1}{2}\int_{-\infty}^{\infty} xe^{-|x|}\,dx = 0$ since $xe^{-|x|}$ is an odd function .

$$\text{variance} = V(X)$$

$$= \int_{-\infty}^{\infty} x^2\, f(x)dx - \mu^2$$

$$= \frac{1}{2}\int_{-\infty}^{\infty} x^2 e^{-|x|}\, dx$$

$$= \frac{1}{2}\int_{0}^{\infty} 2x^2 e^{-x}\, dx = \left[x^2(-e^{-x}) - 2x(e^{-x}) + 2(-e^{-x})\right]_{0}^{\infty} = 2.$$

3. If the probability density function $f(x) = \begin{cases} \dfrac{sinx}{2} & \text{if } 0 \le x \le \pi \\ 0 & \text{otherwise} \end{cases}$.

Find mean,median,mode and $P(0 < x < \frac{\pi}{2})$.

Sol: Mean $= \mu = \int_{-\infty}^{\infty} x\, f(x)dx = \frac{1}{2}\int_{0}^{\pi} x\frac{sinx}{2}dx = \frac{1}{2}[-xcosx + sinx]_{0}^{\pi} = \frac{\pi}{2}$.

Let M be the Median then

$$\int_{0}^{M} f(x)dx = \int_{M}^{\pi} f(x)dx = \frac{1}{2} \ \forall\, x \in (-\infty, \infty)$$

$$\int_{0}^{M} \frac{sinx}{2}dx = \int_{M}^{\pi} \frac{sinx}{2}dx = \frac{1}{2} \ \forall\, x \in (-\infty, \infty)$$

consider $\int_{M}^{\pi} \frac{sinx}{2}dx = \frac{1}{2}$ then $(-cosx)_{M}^{\pi} = 1$

$$\therefore M = \frac{\pi}{2}$$

Since $f(x) = \begin{cases} \dfrac{sinx}{2} & \text{if } 0 \le x \le \pi \\ 0 & \text{otherwise} \end{cases}$

To find maximum, we have $f'(x) = 0$

i.e, $cosx = 0$ implies that $x = \frac{\pi}{2}$

and $f''(x) = -\frac{sinx}{2}$ which is less than 0 at $x = \frac{\pi}{2}$

$\therefore$ Mode $= \frac{\pi}{2}$.

4.If the distributed function is given by

$$F(x) = \begin{cases} 0 \text{ if } x \le 1 \\ k(x-1)^4 \text{ if } 1 \le x \le 3 \\ 1 \text{ if } x > 3 \end{cases}$$

Find k, f(x), mean.

Sol: Cumulative distribution function of $f(x)$ is given by

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

i.e, $f(x) = \begin{cases} 0 \text{ if } x \le 1 \\ 4k(x-1)^3 \text{ if } 1 \le x \le 3 \\ 0 \text{ if } x > 3 \end{cases}$

Since total area under the probability curve is 1 i. e, $\int_a^b f(x)dx = 1$

$$\int_1^3 4k(x-1)^3 dx = 1$$

$$[k(x-1)^4]_1^3 = 1$$

$$\therefore k = \frac{1}{16}$$

$$\therefore f(x) = \begin{cases} \dfrac{1}{4} \begin{array}{l} 0 \text{ if } x \le 1 \\ (x-1)^3 \text{ if } 1 \le x \le 3 \\ 0 \text{ if } x > 3 \end{array} \end{cases}$$

Mean=$\mu = \int_{-\infty}^{\infty} x\, f(x)dx = \frac{1}{4}\int_1^3 x(x-1)^3 dx = 19.6.$

Multiple Random variables

Discrete two dimensional random variable:

Joint probability mass function is defined as $f(x, y) = P(X = x_i, Y = y_i)$

Joint probability distribution function is defined as

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

Marginal probability mass functions of X and Y are defined as

$$P(X = x_i) = p(x_i) = \sum_j p(x_i, y_j)$$

$$P(Y = y_j) = p(y_j) = \sum_i p(x_i, y_j)$$

Continuous two dimensional random variable:

Joint probability density function is defined as

$f_{XY}(x, y) = P(x \le X \le x + dx, y \le Y \le y + dy)$

and $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{XY}(x, y) \, dxdy = 1$

Joint probability distribution function is defined as

$F_{XY}(x, y) = P(X < x_i, Y < y_i) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{XY}(x, y) \, dxdy$

and $f_{XY}(x, y) = \frac{\partial^2}{\partial x \, \partial y} [F_{XY}(x, y)]$

Marginal probability density functions of X is defined as

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

Marginal probability density functions of Y is defined as

$$f_Y(y) = \int_{-\infty}^{\infty} f_{XY}(x, y) \, dx$$

Problems

1.For the following 2-d probability distribution of X and Y

X\Y	1	2	3	4
1	0.1	0	0.2	0.1
2	0.05	0.12	0.08	0.01
3	0.1	0.05	0.1	0.09

Find i) $P(X \leq 2, Y = 2)$ ii) $F_X(2)$ iii)$P(Y=3)$ iv) $P(X < 3, Y \leq 4)$ v) $F_y(3)$.

Sol: Given

X\Y	1	2	3	4
1	0.1	0	0.2	0.1
2	0.05	0.12	0.08	0.01
3	0.1	0.05	0.1	0.09

i)$P(X \leq 2, Y = 2) = P(X = 1, Y = 2) + P(X = 2, Y = 2)$

$\qquad = 0 + 0.12$

$\qquad = 0.12$

ii) $F_X(2) = P(X \leq 2) = P(X = 1) + P(X = 2)$

$\qquad = \Sigma_j\, p(x_i, y_j) + \Sigma_j\, p(x_i, y_j)$

$\qquad = (0.1 + 0 + 0.2 + 0.1) + (0.05 + 0.2 + 0.08 + 0.1)$

$\qquad = 0.66$

iii) $P(Y=3) = \Sigma_i\, p(x_i, y_j)$

$\qquad = 0.2 + 0.08 + 0.1$

$\qquad = 0.38.$

iv) $P(X < 3, Y \leq 4) = P(X < 3, Y = 1) + P(X < 3, Y = 2) + P(X < 3, Y = 3)$

$\qquad\qquad + P(X < 3, Y = 4)$

$\qquad = P(X = 1, Y = 1) + P(X = 2, Y = 1) + P(X = 1, Y = 2)$

$\qquad + P(X = 2, Y = 2) + P(X = 1, Y = 3) + P(X = 2, Y = 3)$

$\qquad + P(X = 1, Y = 4) + P(X = 2, Y = 4)$

$\qquad = (0.1 + 0 + 0.2 + 0.1) + (0.05 + 0.2 + 0.08 + 0.1)$

$\qquad = 0.66$

v)$F_y(3) = P(Y \leq 3) = P(Y = 1) + P(Y = 2) + P(Y = 3)$

$\qquad = (0.1 + 0.05 + 0.1) + (0 + 0.12 + 0.05) + (0.2 + 0.08 + 0.1)$

$\qquad = 0.8$

2.Suppose the random variables X and Y have the joint density function defined by

$$f(x, y) = \begin{cases} c(2x + y) & \text{if } 2 < x < 6, 0 < y < 5 \\ 0 & \text{otherwise} \end{cases}$$

Find i)c   ii)$P(X>3, Y>2)$ iii) $P(X>3)$

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

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

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

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

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

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

ii) $P(X > 3, Y > 2) = \int_3^6 \int_2^5 f(x, y)\,dydx$

$$= \int_3^6 \int_2^5 \frac{1}{210}(2x + y)\,dydx$$

$$= \frac{1}{210} \int_3^6 \left(2xy + \frac{y^2}{2}\right)\Big|_2^5 dx = \frac{15}{28}.$$

iii)) $P(X > 3) = \frac{1}{210} \int_3^6 \int_0^5 f(x, y)\,dydx$

$$= \frac{1}{210} \int_3^6 \int_0^5 (2x + y)\,dydx$$

$$= \frac{1}{210} \int_3^6 \left(2xy + \frac{y^2}{2}\right)\Big|_0^5 dx$$

$$= \frac{1}{210} \int_3^6 \left(10x + \frac{25}{2}\right) dx$$

$$= \frac{1}{210}\left[10x^2 + \left(10x + \frac{25x}{2}\right)\right]_3^6 = \frac{23}{28}.$$

3.The joint density function defined by

$$f(x, y) = \begin{cases} c(xy) \text{ if } 1 < x < 3, 2 < y < 4 \\ \quad 0 \quad \text{otherwise} \end{cases}$$

    Find   i) c

        ii) Marginal probability density functions of X and Y

        iii) Show that X and Y are   independent.

Sol: Since $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)\,dxdy = 1$

$$\int_2^4 \int_1^3 c(xy)\,dxdy = 1$$

$$\int_2^4 cy\left(\frac{x^2}{2}\right)_1^3 dy = 1$$

$$\frac{8c}{2}\left(\frac{y^2}{2}\right)_2^4 = 1 \quad \therefore c = \frac{1}{24}.$$

ii) Marginal probability density functions of X and Y

Marginal probability density functions of X is

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x,y)\, dy = \frac{1}{24}\int_2^4 xy\, dy = \frac{x}{4}$$

Marginal probability density functions of Y is

$$f_Y(y) = \int_{-\infty}^{\infty} f_{XY}(x,y)\, dx = \frac{1}{24}\int_1^4 xy\, dx = \frac{y}{6}$$

iii) Clearly $f_{XY}(x,y) = \frac{xy}{24} == \frac{x}{4}\frac{y}{6} = f_X(x)\, f_Y(y)$

Therefore, X and Y are independent.

Conditional probability density function :

Conditional probability density function of X on Y is

$$f_{XY}(X/Y) = \frac{f_{XY}(x,y)}{f_Y(y)}$$

Conditional probability density function of Y on X is

$$f_{XY}(Y/X) = \frac{f_{XY}(x,y)}{f_X(X)}$$

4.The joint density function defined by

$$f(x,y) = \begin{cases} \left(x^2 + \frac{xy}{3}\right) & \text{if } 0 < x < 1, 0 < y < 2 \\ 0 & \text{otherwise} \end{cases}$$

Find

i) **Conditional probability density functions.**
ii) **Marginal probability density functions**
iii) **Check whether the functions X and Y are independent or not**

Sol: Given $f(x,y) = \begin{cases} \left(x^2 + \frac{xy}{3}\right) & \text{if } 0 < x < 1, 0 < y < 2 \\ 0 & \text{otherwise} \end{cases}$

Marginal probability density functions of X is

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x,y)\, dy = \int_0^2 \left(x^2 + \frac{xy}{3}\right) dy = 2x\left(x + \frac{1}{3}\right)$$

Marginal probability density functions of Y is

$$f_Y(y) = \int_{-\infty}^{\infty} f_{XY}(x,y)\, dx = \int_0^1 \left(x^2 + \frac{xy}{3}\right) dx = \frac{1}{3} + \frac{y}{6}$$

Here $\quad f_Y(y) \; f_X(x) = 2x(x + \frac{1}{3})(\frac{1}{3} + \frac{y}{6})$

Therefore, $\quad f_{XY}(x, y) \neq f_X(x) \; f_Y(y)$

Hence X and Y are not Independent.

Conditional probability density function of X on Y is

$$f_{XY}(X/Y) = \frac{f_{XY}(x,y)}{f_Y(y)} = \frac{(x^2 + \frac{xy}{3})}{(\frac{1}{3} + \frac{y}{6})}$$

Conditional probability density function of Y on X is

$$f_{XY}(Y/X) = \frac{f_{XY}(x,y)}{f_X(X)} = \frac{(x^2 + \frac{xy}{3})}{2x(x + \frac{1}{3})}$$

TUTORIAL QUESTIONS

1) For the following bivariate, find, (i) P(X≤2, Y=2)
   (ii) $F_X(2)$
   (iii) P(Y=3)
   (iv) P(X<3, Y≤4)   and (v) $F_Y(3)$

X/Y	1	2	3	4
1	0.1	0	0.2	0.1
2	0.05	0.12	0.08	0.01
3	0.1	0.05	0.1	0.09

2a) A random variable X has following P.D. Find (i) k (ii) P(X>6) (iii) Find 'c' if $P(X \leq c) > \frac{1}{2}$

X	0	1	2	3	4	5	6	7
P(X)	0	k	2k	2k	3k	k^2	$2k^2$	$7k^2+k$

b) If F(x) is the distribution function of X given by

$$F(x) = \begin{cases} 0, x \leq 1 \\ k(x-1)^4, 1 < x \leq 3 \\ 1, x > 3 \end{cases} \text{ Find (i) f(x) (ii) k (iii) mean}$$

3) The joint density function given by $f_{xy}(x,y) = \begin{cases} cxy, 1 < x < 3, 2 < y < 4 \\ 0, otherwise \end{cases}$

   Find (i) c (ii) marginal density functions (iii) conditional density functions
   (iv) S.T X ,Y are independent

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

   b) If the p.d.f is $f(x) = \frac{K}{1+x^2}$, $-\infty < x < \infty$ Find (i)K (ii) Probability distribution function

5a) A continuous r.v has the p.d.f $f(x) = \begin{cases} kxe^{-\alpha x}, x \geq 0, \alpha > 0 \\ 0, elsewhere \end{cases}$ Determine (i)K, (ii) mean

 (iii) variance

   b) A random sample with replacement of size 2 is taken from S ={1,2,3}. Let X denote sum
of 2 no. s taken .(i)Write probability distribution (ii) find mean

ASSIGNMENT QUESTIONS

1. If the p.d.f of a r.v x is given by $f(x) = \begin{cases} k(1-x^2), & 0 < x < 1 \\ 0, & otherwise \end{cases}$

   find i) k

2. If F(x) is the distribution function of x is given by $F(X) = \begin{cases} 0 & if \ x \leq 1 \\ k(x-1)^4 & if \ 1 < x \leq 3 \\ 1 & if \ x > 3 \end{cases}$

   Determine i) f(x) ii) k iii)mean

3. If X is a continuous r.v and Y=kX+c prove that E(Y)=kE(X)+y and V(Y)=$k^2 V(X)$, where V stands for Variance

4. A random variable has the following probability function

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

   Determine i) K   ii) Mean iii)Variance iv) P(2< X < 6).

5. The joint density function given by $f_{xy}(x,y) = \begin{cases} cxy, 1 < x < 3, 2 < y < 4 \\ 0, otherwise \end{cases}$

   Find (i) c (ii) marginal density functions (iii) conditional density functions
   (iv) S.T X ,Y are independent

UNIT 2
PROBABILITY DISTRIBUTIONS

OBJECTIVE

To learn some of the important probability distributions like

Binomial, Poisson Distributions (discrete case) and the Normal

Distribution (continuous case).

OUTCOME

Provide very good insight which is essential for industrial applications

by learning probability distributions.

PROBABILITY DISTRIBUTIONS

Binomial Distribution: A Random variable 'X' has binomial distribution if it assumes only non-negative values with probability mass function given by

$$p(x = r) = \begin{cases} n_{c_r} p^r q^{n-r} & r = 0,1,2, ---- n \\ 0 & otherwise \end{cases}$$

$$= b(r; n, p)$$

Conditions For Applicability Of Binomial Distributions:

1. Number of trials must be finite (n is finite)
2. The trails are independent
3. There are only two possible outcomes in any event i.e., success and failure.
4. Probability of success in each trail remains constant.

Examples:

1. Tossing a coin n times
2. Throwing a die
3. No. of defective items in the box

Mean Of The Binomial Distribution

$$\mu = \sum_{r=0}^{n} r. P(r)$$

$$= \sum_{r=0}^{n} r. n_{c_r} P^r q^{n-r}$$

$$= n_{c_1} p^1 q^{n-1} + 2n_2 P^r q^{n-2} + 3n_{c_3} p^3 q^{n-3} + \cdots \ldots. nn_{c_n} p^n q^{n-n}$$

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

$$= np[q^{(n-1)} + (n-1)_{c_1} p^1 q^{(n-1)-1} + ---- + p^{n-1}]$$

$$= np[p + q]^{n-1}$$

$$= np \qquad \text{since } [p + q = 1]$$

Mean = np.

Variance Of The Binomial Distribution

$$\sigma^2 = \sum_{r=0}^{n} r^2 p(r) - \mu^2$$

$$= \sum_{r=0}^{n} [r(r-1) + r]P(r) - \mu^2$$

$$= \sum_{r=0}^{n} r(r-1)P(r) + \sum_{r=0}^{n} r.P(r) - n^2 p^2$$

$$= \sum_{r=0}^{n} r(r-1)n_{c_r}P^r q^{n-r} + np - n^2 P^2$$

let $\sum_{r=0}^{n} r(r-1)P(r) = \sum_{r=0}^{n} r(r-1)n_{c_r}P^r q^{n-r} = 2\, n_{c_{2r}}P^2 q^2 n^{n-2} +$

$$6n_{c_3}P^3 q^{n-3} + 12n_{c_r}P^4 q^{n-4} + \cdots - + n(n-1)\,P^n$$

$$= n(n-1)P^2 \left[q^{n-2} + +(n-2)_{c_1}p^1 q^{(n-2)-1} + - - - + p^2 \right]$$

$$= n(n-1)P^2 (p+q)^{n-2}$$

$$= n^2 P^2 - nP^2$$

$$\sigma^2 = n^2 P^2 - nP^2 + np - n^2 P^2$$

$$= np(1-p)$$

$$= npq.$$

Problems

1.In tossing a coin 10 times simultaneously. Find the probability of getting

i)at least 7 heads ii) almost 3 heads iii)exactly 6 heads.

Sol: Given $n = 10$

Probability of getting a head in tossing a coin $= \frac{1}{2} = p$.

Probability of getting no head $= q = 1 - \frac{1}{2} = \frac{1}{2}$.

The probability of getting r heads in a throw of 10 coins is
$$P(X = r) = p(r) = 10_{C_r} \left(\frac{1}{2}\right)^r \left(\frac{1}{2}\right)^{10-r}; r = 0,1,2,\ldots\ldots,10$$

(i)   Probability of getting at least seven heads is given by
$$P(X \geq 7) = P(X = 7) + P(X = 8) + P(X = 9) + P(X = 10)$$
$$= 10_{C_7}\left(\frac{1}{2}\right)^7 \left(\frac{1}{2}\right)^{10-7} + 10_{C_8}\left(\frac{1}{2}\right)^8 \left(\frac{1}{2}\right)^{10-8} + 10_{C_9}\left(\frac{1}{2}\right)^9 \left(\frac{1}{2}\right)^{10-9} + 10_{C_{10}}\left(\frac{1}{2}\right)^{10}$$

$$= \frac{1}{2^{10}}\left[10_{C_7} + 10_{C_8} + 10_{C_9} + 10_{C_{10}}\right]$$

$$= \frac{1}{2^{10}}\left[120 + 45 + 10 + 1\right]$$

$$= \frac{176}{1024}$$

$$= 0.1719$$

ii) Probability of getting at most 3 heads is given by

$$P(X \le 3) = P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3)$$

$$= 10_{c_1} \left(\frac{1}{2}\right)^1 \left(\frac{1}{2}\right)^{10-1} + 10_{c_2} \left(\frac{1}{2}\right)^2 \left(\frac{1}{2}\right)^{10-2} + 10_{c_3} \left(\frac{1}{2}\right)^3 \left(\frac{1}{2}\right)^{10-3} + 10_{c_0} \left(\frac{1}{2}\right)^{10}$$

$$= \frac{1}{2^{10}} \left[10_{c_0} + 10_{c_1} + 10_{c_2} + 10_{c_3}\right]$$

$$= \frac{1}{2^{10}} [120 + 45 + 10 + 1]$$

$$= \frac{176}{1024}$$

$$= 0.1719$$

iii) Probability of getting exactly six heads is given by

$$P(X = 6) = 10_{c_6} \left(\frac{1}{2}\right)^6 \left(\frac{1}{2}\right)^{10-6}$$

$$= 0.205.$$

2.In 256 sets of 12 tosses of a coin ,in how many cases one can expect 8 Heads and 4 Tails.

Sol: The probability of getting a head, $p = \frac{1}{2}$

The probability of getting a tail, $q = \frac{1}{2}$

Here $n = 12$

The probability of getting 8heads and 4Tails in 12trials $= P(X = 8) = 12_{C_8} \left(\frac{1}{2}\right)^8 \left(\frac{1}{2}\right)^4$

$$= \frac{12!}{8!\,4!} \left(\frac{1}{2}\right)^{12} = \frac{495}{2^{12}}$$

The expected number of getting 8 heads and 4 Tails in 12 trials of such cases in 256 sets

$$= 256 \times P(X = 8) = 2^8 \times \frac{495}{2^{12}} = \frac{495}{16} = 30.9375 \sim 31$$

3.Find the probability of getting an even number 3 or 4 or 5 times in throwing a die 10 times

Sol: Probability of getting an even number in throwing a die $= \frac{3}{6} = \frac{1}{2} = p.$

Probability of getting an odd number in throwing a die $=q=\frac{1}{2}$.

∴Probability of getting an even number 3 or 4 or 5 times in throwing a die 10 times is

$$P(X=3)+P(X=4)+P(X=5)$$

$$=10_{C_3}\left(\frac{1}{2}\right)^3\left(\frac{1}{2}\right)^{10-3}+10_{C_4}\left(\frac{1}{2}\right)^4\left(\frac{1}{2}\right)^{10-4}+10_{C_5}\left(\frac{1}{2}\right)^5\left(\frac{1}{2}\right)^{10-5}$$

$$=\frac{1}{2^{10}}\left[10_{C_3}+10_{C_4}+10_{C_5}\right]$$

$$=\frac{1}{2^{10}}[120+252+210]$$

=0.568.

4.Out of 800 families with 4 children each ,how many could you expect to have

a)three boys b)five girls c) 2 or 3 boys d)at least 1 boy.

Sol: : Given $n=5, N=800$

Let having boys be success

Probability of having a boy $=\frac{1}{2}=p$.

Probability of having girl $=q=1-\frac{1}{2}=\frac{1}{2}$.

The probability of having r boyss in 5 children is
$$P(X=r)=p(r)=5_{C_r}\left(\frac{1}{2}\right)^r\left(\frac{1}{2}\right)^{5-r};r=0,1,2\ldots\ldots5$$

a)Probability of having 3 boys is given by
$$P(X=3)=5_{C_r}\left(\frac{1}{2}\right)^3\left(\frac{1}{2}\right)^{5-3}=\frac{5}{16}$$

Expected number of families having 3 boys $=N\,p(3)=800(\frac{5}{16})=250$ families.

b) Probability of having 5 girls $=$ Probability of having no boys is given by
$$P(X=0)=5_{C_0}\left(\frac{1}{2}\right)^0\left(\frac{1}{2}\right)^{5-0}=\frac{1}{32}$$

Expected number of families having 5 girls $=N\,p(0)=800(\frac{1}{32})=25$ families.

c) Probability of having either 2 or 3 boys is given by
$$P(X=2)+P(X=3)=5_{C_2}\left(\frac{1}{2}\right)^2\left(\frac{1}{2}\right)^{5-2}+5_{C_3}\left(\frac{1}{2}\right)^3\left(\frac{1}{2}\right)^{5-3}=\frac{5}{18}$$

Expected number of families having 3 boys $=N\,p(3)=800(\frac{5}{8})=500$ families.

d) Probability of having at least 1 boy is given by

$$P(X\geq1)=1-P(X=0)$$

$$= 1 - 5_{C_0} \left(\frac{1}{2}\right)^0 \left(\frac{1}{2}\right)^{5-0} = \frac{31}{32}$$

Expected number of families having at least 1 boy $= 800(\frac{31}{32}) = 775$ families.

5.Fit a Binomial distribution for the following data.

x	0	1	2	3	4	5
f	2	14	20	34	22	8

Sol: Given n= 5,$\sum f = 2 + 14 + 20 + 34 + 22 + 8 = 100$

$\sum x_i f_i == 0(2) + 1(14) + 2(20) + 3(34) + 4(22) + 5(8) = 284$

$\therefore$ Mean of the distribution $= \frac{\sum x_i f_i}{\sum f_i} = \frac{284}{100} = 2.84$

We have Mean of the binomial distribution $= np = 2.84$

$$\therefore p = \frac{2.84}{5} = 0.568; q = 1 - 0.568 = 0.432.$$

Table To Fit Binomial Distribution

X	$P(x_i)$	$E(x_i)$
0	$5_{C_0}(0.568)^0 (0.432)^{5-0} = 0.02$	N p(0) = 100(0.02)=2
1	$5_{C_1}(0.568)^1 (0.432)^{5-1} = 0.09$	9
2	$5_{C_2}(0.568)^2 (0.432)^{5-2} = 0.26$	26
3	$5_{C_3}(0.568)^3 (0.432)^{5-3} = 0.34$	34
4	$5_{C_4}(0.568)^4 (0.432)^{5-4} = 0.22$	22
5	$5_{C_5}(0.568)^5 (0.432)^{5-5} = 0.059$	5.9

Fitted Binomial distribution is

x	0	1	2	3	4	5
f	2	10	26	34	22	6

Recurrence Relation

$$p(r+1) = n_{C_{r+1}}(p)^{r+1}(q)^{n-r-1} \ldots\ldots\ldots(1)$$

$$p(r) = n_{C_r}(p)^r(q)^{n-r}\ldots\ldots\ldots\ldots\ldots\ldots\ldots(2)$$

$$\frac{(1)}{(2)} = \frac{p(r+1)}{p(r)} = \frac{n_{C_{r+1}}(p)^{r+1}(q)^{n-r-1}}{n_{C_r}(p)^r(q)^{n-r}}$$

$$\therefore \frac{p(r+1)}{p(r)} = \frac{n_{C_{r+1}}}{n_{C_r}}\left(\frac{p}{q}\right)$$

$$p(r+1) = \frac{n_{C_{r+1}}}{n_{C_r}}\left(\frac{p}{q}\right)p(r).$$

Poisson Distribution

A random variable 'X' follows Poisson distribution if it assumes only non-negative values with probability mass function is given by

$$P(x = r) = P(r, \lambda) = \begin{cases} \dfrac{e^{-\lambda}\lambda^r}{r!} & for \quad y = 0,1,-- \ (\lambda > 0) \\ 0 & otherwise \end{cases}$$

Conditions For Poisson Distribution

1. The number of trials are very large (infinite)
2. The probability of occurrence of an event is very small $(\lambda = np)$
3. $\lambda = np = finite$

Examples:

1. The number of printing mistakes per page in a large text
2. The number of telephone calls per minute at a switch board
3. The number of defective items manufactured by a company.

Recurrence Relation

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

$$P(r) = \frac{e^{-\lambda}\lambda^r}{(r)!} \quad -------(2)$$

$$\frac{1}{2} = \frac{P(r + 1)}{P(r)} = \frac{e^{-\lambda}\lambda^2.\lambda}{(r + 1)r!} X \frac{r!}{e^{-\lambda}\lambda^2}$$

$$P(r + 1) = \left(\frac{\lambda}{r + 1}\right)P(r) \ for \ r = 0,1,2 ----$$

Problems

1.Using Recurrence relation find probability when x=0,1,2,3,4,5, if mean of P.D is 3.

Sol: We have
$$P(r + 1) = \left(\frac{\lambda}{r+1}\right)P(r) \ for \ r = 0,1,2 ----(1)$$

Given λ= 3

$$P(0) = \frac{e^{-3}\lambda^0}{(0)!} = e^{-3} \ \text{[by definition of Poisson distribution]}$$

From (1),

For $r = 0$, $P(1) = \left(\frac{3}{0+1}\right)P(0)$ =$3\ e^{-3}$

For $r = 1$, $P(2) = \left(\frac{3}{1+1}\right)P(0) = \frac{3}{2}\,e^{-3}$

For $r = 2$, $P(3) = \left(\frac{3}{2+1}\right)P(0) = e^{-3}$

F For $r = 3$, $P(4) = \left(\frac{3}{3+1}\right)P(0) = \frac{3}{4}e^{-3}$

or $r = 4$, $P(5) = \left(\frac{3}{4+1}\right)P(0) = \frac{3}{5}\,e^{-3}$.

2.If X is a random variable such that $3P(X = 4) = \frac{P(X=2)}{2} + P(X = 0)$.

Find mean, $P(X \leq 2)$.

Sol: Given $3P(X = 4) = \frac{P(X=2)}{2} + P(X = 0)$…..(1)

Since X is a Poisson variable,

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

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

Solving it we get $\lambda^4 - 2\lambda^2 - 4 = 0$

Taking $\lambda^2 = k$, we get $k^2 - 2k - 4 = 0$

$$\therefore k = 4, -2$$

$$\therefore \lambda^2 = 4$$

$$\text{i.e } \lambda = 2$$

Therefore, Mean of the Poisson distribution = 2

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

$$= \frac{e^{-2}\lambda^0}{0!} + \frac{e^{-2}2^1}{1!} + \frac{e^{-\lambda}2^2}{2!} = 0.54.$$

3.A car hire firm has 2 cars which it hires out day by day.The number of demands for a car on each day is distributed as poisson with mean 1.5 Calculate the proportion of days

i)on which there is no demand

ii) on which demand is refused.

Sol: Let number of demands for cars be the success.

Given mean = 1.5= λ

Using Poisson distribution,

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

i)Probability that there is no demand for car is

$$P(x = 0) = \frac{e^{-1.5}(1.5)^0}{0!} = 0.223$$

Expected number of days that there is no demand $= NP(0) = 365(0.223) = 81.39 \sim 81 \ days$

ii) Probability that demand refused for car is

$$P(x > 2) = 1 - P(x = 0) - P(x = 1) - P(x = 2)$$

$$= 1 - \frac{e^{-1.5}(1.5)^0}{0!} - \frac{e^{-1.5}(1.5)^1}{1!} - \frac{e^{-1.5}(1.5)^2}{2!} = 0.191$$

Expected number of days that demand refused for car $= NP(x > 2)$

$$= 365(0.191) = 69.7 \sim 70 \ days.$$

4.The distribution of typing mistakes committed by typist is given below.

Fit a Poisson distribution for it.

Mistakes per page	0	1	2	3	4	5
Number of pages	142	156	69	27	5	1

Sol: Given n= 5,$\sum f = 142 + 156 + 69 + 27 + 5 + 1 = 400$

$\sum x_i f_i == 0(142) + 1(156) + 2(69) + 3(27) + 4(5) + 5(1) = 400$

$$\therefore \text{Mean of the distribution} = \frac{\sum x_i f_i}{\sum f_i}$$

$$= \frac{400}{400} = 1.$$

We have Mean of the Poisson distribution $= \lambda = 1$

Table To Fit Poisson Distribution

X	$P(x_i)$	$E(x_i)$
0	$\frac{e^{-1}(1)^0}{0!}$=0.368	N p(0) =400(0.368)=147.2~147
1	$\frac{e^{-1}(1)^1}{1!}$=0.368	147
2	$\frac{e^{-1}(1)^2}{2!}$=0.184	74
3	$\frac{e^{-1}(1)^3}{3!}$=0.061	24
4		6

5	$\dfrac{e^{-1}(1)^4}{4!} = 0.015$ $\dfrac{e^{-1}(1)^5}{5!} = 0.003$	1

Fitted Poisson distribution is

Mistakes per page	0	1	2	3	4	5
Number of pages	147	147	74	24	6	1

Normal Distribution (Gaussian Distribution)

Let X be a continuous random variable, then it is said to follow normal distribution if its pdf is given by

$$f(x, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \qquad\qquad -\infty \leq x \leq \infty, \mu, \sigma > 0$$

Here , σ are the mean & S.D of X.

Properties Of Normal Distribution

1. Normal curve is always centered at mean
2. Mean, median and mode coincide (i.e., equal)
3. It is unimodal
4. It is a symmetrical curve and bell shaped curve
5. X-axis is an asymptote to the normal curve
6. The total area under the normal curve from $-\infty$ to ∞ is "1"
7. The points of inflection of the normal curve are $\mu \pm \sigma, \mu \pm 3\sigma$
8. The area of the normal curve between

$$\mu - \sigma \text{ to } \mu + \sigma = 68.27\%$$

$$\mu - 2\sigma \text{ to } \mu + 2\sigma = 95.44\%$$

$$\mu - 3\sigma \text{ to } \mu + 3\sigma = 99.73\%$$

9.The curve for normal distribution is given below
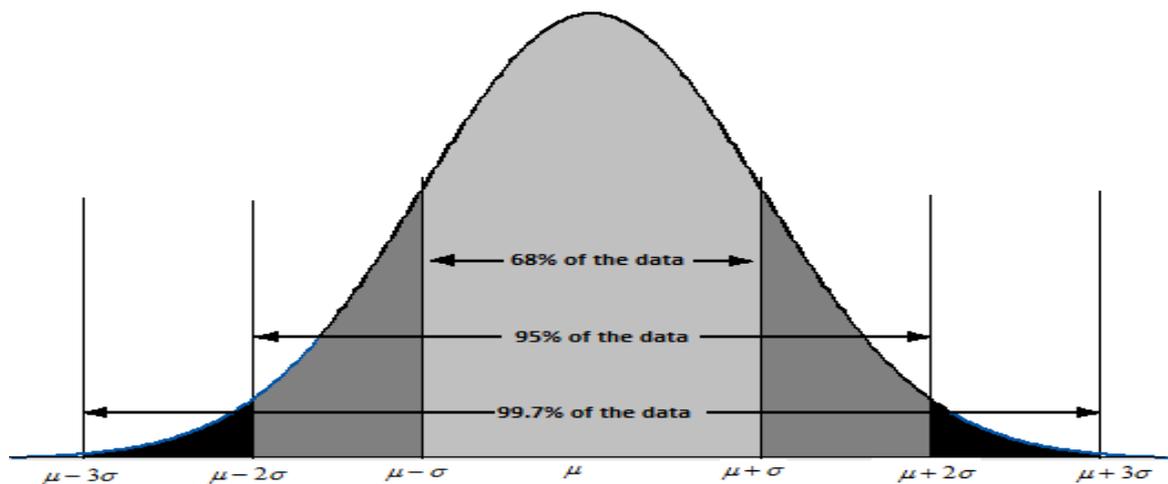
Standard Normal Variable

Let $Z = \frac{x-\mu}{\sigma}$ with mean '0' and variance is '1' then the normal variable is said to be standard normal variable.

Standard Normal Distribution

The normal distribution with man '0' and variance '1' is said to be standard normal distribution of its probability density function is defined by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-u}{\sigma}\right)^2} \qquad -\infty < x \leq \infty$$

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \qquad -\infty \leq x \leq \infty \ (\mu = 0, \sigma = 1)$$

Mean Of The Normal Distribution

Consider Normal distribution with b,σ as parameters Then

$$f(x; b, \sigma) \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-b)^2}{2\sigma^2}}$$

Mean of the continuous distribution is given by

$$\mu = \int_{-\infty}^{\infty} x\, f(x)dx = \int_{-\infty}^{\infty} x\, \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-b)^2}{2\sigma^2}} dx$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} (\sigma z + b)\, e^{-\frac{(z)^2}{2}} dz \quad [\text{Putting } z = \frac{x-b}{\sigma} \text{ so that } dx = \sigma\, dz]$$

$$= \frac{\sigma}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z\, e^{-\frac{(z)^2}{2}} dz + \frac{b}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{(z)^2}{2}} dz$$

$$= \frac{2b}{\sqrt{2\pi}} \int_{-0}^{\infty} e^{-\frac{(z)^2}{2}} dz$$

[since $z\, e^{-\frac{(z)^2}{2}}$ is an odd function and $e^{-\frac{(z)^2}{2}}$ is an even function]

$$\mu = \frac{2b}{\sqrt{2\pi}} \sqrt{\frac{\pi}{2}} = b$$

$$\therefore \text{Mean} = b$$

Variance Of The Normal Distribution

Variance $= \int_{-\infty}^{\infty} x^2 f(x)dx - \mu^2$

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

Let $z = \frac{x-\mu}{\sigma} \implies dx = \sigma dz$

$$= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} (\mu^2 + \sigma^2 z^2 + 2\mu\sigma z)\, e^{-\frac{z^2}{2}} \sigma dz - \mu^2$$

$$= \frac{\mu^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{z^2}{2}} \, dz + \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z^2 \, e^{-\frac{3^2}{2}} \, dz + \frac{2\mu\sigma}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z^2 \, e^{-\frac{3^2}{2}} \, dz - \mu^2$$

$$= \frac{2\mu^2}{\sqrt{2\pi}} \int_{0}^{\infty} e^{-\frac{z^2}{2}} \, dz + \frac{2\sigma^2}{\sqrt{2\pi}} \int_{0}^{\infty} z^2 \, e^{-\frac{3^2}{2}} \, dz - \mu^2$$

$$= \frac{2\sigma^2}{\sqrt{2\pi}} \int_{0}^{\infty} z^2 e^{-\frac{z^2}{2}} \, dz$$

$$\because \frac{z^2}{2} = + \Rightarrow \frac{2z\,dz}{2} = dt \qquad\qquad dz = \frac{dt}{\sqrt{2t}}$$

$$= \frac{2\sigma^2}{\sqrt{2\pi}} \int_{0}^{\infty} (2+)^2 e^t \frac{dt}{\sqrt{2t}}$$

$$= \frac{2\sigma^2}{\sqrt{\pi}} \int_{0}^{\infty} e^{-t} + {}^{\frac{3}{2}-1} . dt$$

$$= \frac{2\sigma^2}{\sqrt{\pi}} \Gamma\left(\tfrac{3}{2}\right)$$

$$= \frac{2\sigma^2}{\sqrt{\pi}} \frac{1}{2} \Gamma\left(\frac{1}{2}\right)$$

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

$$\therefore \sigma^2 = \int_{-\infty}^{\infty} x^2 \, f(x)dx - \mu^2.$$

Median Of The Normal Distribution

Let x=M be the median, then

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

Let $\mu \epsilon (-\infty, M)$

Let $\int_{-\infty}^{\infty} f(x)dx = \int_{-\infty}^{\mu} f(x)dx + \int_{\mu}^{M} f(x)dx = \frac{1}{2}$

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

Let $z = \frac{x-\mu}{\sigma} \Rightarrow dx = \sigma dz$ $\qquad\qquad$ [$\because$ Limits of $z - \infty \to 0$]

$\int_{-\infty}^{\mu} f(x)dx = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{0} e^{-\frac{z^2}{2}} \sigma dz$

$$= \frac{1}{\sqrt{2\pi}} \int_{\infty}^{0} e^{-\frac{t^2}{2}} \, (dt) \text{ (by taking z=-t again)}$$

$$= \frac{1}{\sqrt{2\pi}} \sqrt{\frac{\pi}{2}} = \frac{1}{2}$$

$$\therefore \int_{\mu}^{\mu} f(x) dx = 0 \Rightarrow \mu = M$$

Mode Of The Normal Distribution

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

$$f'(x) = 0 \Rightarrow \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \left(\frac{-1}{2}\right) 2 \left(\frac{x-\mu}{\sigma}\right)\frac{1}{\sigma} = 0$$

$$\Rightarrow x - \mu = 0 \Rightarrow x = \mu$$

$$\Rightarrow x = \mu$$

$$f^{11}(x) = \frac{-1}{\sigma^3\sqrt{2\pi}}\left[e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}.1 + (x-\mu)e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}\left(\frac{-1}{2}\right)2\left(\frac{x-\mu}{\sigma}\right)\frac{1}{\sigma}\right]$$

$$= \frac{-1}{\sigma^3\sqrt{2\pi}} \ [e^0 + 0]$$

$$= \frac{-1}{\sigma^3\sqrt{2\pi}} < 0$$

$\because x = \mu$ is the mode of normal distribution.

Problems :

1.If X is a normal variate, find the area A

  i)   to the left of $z = 1.78$

  ii)   to the right of $z = -1.45$

  iii)   Corresponding to $-0.8 \leq z \leq 1.53$
  iv)   to the left of $z = -2.52$ and to the right of $z = 1.83$.

Sol: i) $P(z < -1.78) = 0.5 - P(-1.78 < z < 0)$

$$= 0.5 - P(0 < z < 1.78)$$

$$= 0.5\text{-}0.4625 = 0.0375.$$

ii) $P(z > -1.45) = 0.5 + P(-1.45 < z < 0)$

$$= 0.5 + P(0 < z < 1.45)$$

$$= 0.5+0.4625 =0.9265.$$

iii) $P(-0.8 \leq z \leq 1.53) = P(-0.8 \leq z \leq 0) + P(0 \leq z \leq 1.53)$

$$= 0.2881+0.4370=0.7251.$$

iv) $P(z < -2.52) = 0.5 - P(0 < z < 2.52)=0.0059$

$$P(z > 1.83) \quad = 0.5 - P(0 < z < 1.83)$$

$$=0.036$$

2.If the masses of 300 students are normally distributed with mean 68 kgs and standard deviation 3kgs.How many students have masses

i)greater than 72kgs.

ii)less than or equal to 64 kgs

iii)between 65 and 71 kgs inclusive.

Sol: Given N=300,$\mu = 68, \sigma = 3$.Let X be the masses of the students.

i)   Standard normal variate for X=72 is

$$z = \frac{x - \mu}{\sigma} = \frac{72 - 68}{3} = 1.33$$

$P(X > 72)= P(z > 1.33)$

$$== 0.5 - P(0 < z < 1.33)$$

$$=0.5\text{-}0.4082$$

$$=0.092$$

Expected number of students greater than 72 = E(X>72)
$$=300(0.092)$$
$$=27.54\text{~}28 \text{ students}$$

ii) Standard normal variate for X=64 is

$$z = \frac{x - \mu}{\sigma} = \frac{64 - 68}{3} = -1.33$$

$P(X \leq 64)= P(z \leq -1.33)$

$$== 0.5 - P(0 < z < 1.33) \text{ (Using symmetry)}$$

$$=0.5\text{-}0.4082$$

$$=0.092$$

Expected number of students less than or equal to 64 = E(X less than or equal to 64)

$$=300(0.092)$$
$$=27.54\sim28 \text{ students}.$$

iii)Standard normal variate for X=65 is

$$z_1 = \frac{x - \mu}{\sigma} = \frac{65 - 68}{3} = -1$$

Standard normal variate for X=71 is

$$z_2 = \frac{x - \mu}{\sigma} = \frac{71 - 68}{3} = 1$$

$$P(65 \leq X \leq 71) = P(-1 \leq z \leq 1)$$
$$= P(-1 \leq z \leq 0) + P(-0 \leq z \leq 1)$$
$$= 2 P(-0 \leq z \leq 1)$$

$$=2(0.341)= 0.6826$$

$$E(65 \leq X \leq 71) = 300(0.6826) = 205 \text{ } Students.$$

∴Expected number of students between 65 and 71 kgs inclusive = 205 students.

3.In a normal distribution 31% of the items are under 45 and 8% of the items are over 64. Find mean and variance of the distribution.

Sol: Given $P(X < 45)$= 31% = 0.31

And $P(X > 64)$= 8% = 0.08

Let Mean and variances of the normal distributions are μ, σ^2.

Standard normal variate for X is

$$z = \frac{x - \mu}{\sigma}$$

Standard normal variate for X_1=45 is

$$z_1 = \frac{X_1 - \mu}{\sigma} = \frac{45 - \mu}{\sigma}$$
$$\Rightarrow \mu + \sigma z_1 = 45 \ldots \ldots \ldots (1)$$

Standard normal variate for X_2=64 is

$$z_2 = \frac{X_2 - \mu}{\sigma} = \frac{64 - \mu}{\sigma}$$

$$\Rightarrow \mu + \sigma z_2 = 64 \ldots \ldots \ldots (2)$$

From normal curve ,we have $P(-z_1 \leq z \leq 0) = 0.19$

$$\Rightarrow z_1 = -0.5$$

$P(0 \leq z \leq z_2) = 0.42$

$$\Rightarrow z_2 = 1.41$$

substituting the values of z_1, z_2 in (1) and (2),we get

$\mu = 50, \sigma^2 = 98$.

4. In a normal distribution 7% of the items are under 35 and 89% of the items are under 63. Find mean and variance of the distribution.

Sol: Given $P(X < 35) = 7\% = 0.07$

And $P(X < 63) = 89\% = 0.89$

Let Mean and variances of the normal distributions are μ, σ^2.

Standard normal variate for X is

$$z = \frac{x - \mu}{\sigma}$$

Standard normal variate for $X_1 = 35$ is

$$z_1 = \frac{X_1 - \mu}{\sigma} = \frac{35 - \mu}{\sigma}$$
$$\Rightarrow \mu + \sigma z_1 = 35 \ldots \ldots \ldots (1)$$

Standard normal variate for $X_2 = 63$ is

$$z_2 = \frac{X_2 - \mu}{\sigma} = \frac{63 - \mu}{\sigma}$$
$$\Rightarrow \mu + \sigma z_2 = 63 \ldots \ldots \ldots (2)$$

Given $P(X < 35) = P(z < z_1)$

$\quad 0.07 = 0.5 - P(-z_1 \leq z \leq 0)$

$\quad P(0 \leq z \leq z_1) = 0.43$

From normal curve ,we *have*

$$\Rightarrow z_1 = 1.48$$

We have $P(X < 63) = P(z < z_2)$

$\quad 0.89 = 0.5 + P(0 \leq z \leq z_2)$

$\quad P(0 \leq z \leq z_2) = 0.39$

From normal curve ,we *have*

$$\Rightarrow z_2 = 1.23$$

substituting the values of z_1, z_2 in (1) and (2),we get

$\mu = 50, \sigma^2 = 100$.

TUTORIAL QUESTIONS

1. Out of 800 families with 4 children each, how many could you expect to have

   i)three boys ii)five girls iii) 2 or 3 boys iv) at least 1 boy.

2. Suppose 2% of the people on the average are physically handicapped .Find

   i) the probability of   finding 3 or more Physically handicapped

   ii) the probability of finding atmost 2 are handicapped

3. If X is a normal variate with mean 30 and variance 25.Find
   i)P(26≤X≤40) ii)P(X≥45).

4. In a normal distribution, 7% of the items are under 35 and 89% are under 63.
   Determine the mean and variance of the distribution.

5.Fit a Poisson distribution to the following data

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

ASSIGNMENT QUESTIONS

1. A sales tax officer has reported that the average sales of the 500 business that he has to deal with during a year is Rs.36,000 with a standard deviation of Rs.10,000. Assuming that the sales in these business are normally distributed ,find :

   i) The number of business as the sales of which are greater than Rs.40,000
   ii) The percentage of business the sales of which are likely to range between Rs.30,000 and Rs.40,000

2. If the variance of a Poisson variate is 3 . Find the probability that
   i) $P(x=0)$ ii)$P(1\leq x\leq 4)$ iii) $P(x>2)$

3. The mean and SD of a normal variate are 8 and 4 respectively .Find i) $P(5\leq x\leq 10)$
   ii)$P(x\geq 5)$

4. Average number of accidents on any day on a national highway is 1.8 .determine the probability that the number of accidents are i) atleast one ii) at the most one iii) exactly one

5.If X is a normal variate , find

i) to the left of Z=-1.78
ii) to the right of Z=-1.45
iii) corresponding to -0.80$\leq$Z$\leq$1.53
iv) to the left of Z=-2.52 and to the right of Z=1.83
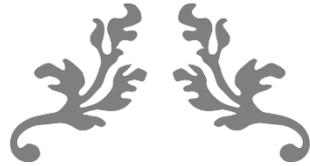
UNIT 3

CORRELATION

AND

REGRESSION

OBJECTIVE

To understand linear relationship between two variables and also to predict

how a dependent variable changes based on adjustments to an independent variable.

OUTCOME

Make data-driven decisions by using correlation and regression.

<h1 style="text-align:center">CORRELATION AND REGRESSION</h1>

CORRELATION

Introduction

In a bivariate distribution and multivariate distribution we may be interested to find if there is any relationship between the two variables under study. Correlation refers to the relationship between two or more variables. The correlation expresses the relationship or interdependence of two sets of variables upon each other.

Definition Correlation is a statistical tool which studies the relationship b/w 2 variables & correlation analysis involves various methods & techniques used for studying & measuring the extent of the relationship b/w them.

Two variables are said to be correlated if the change in one variable results in a corresponding change in the other.

The Types of Correlation

1) Positive and Negative Correlation: If the values of the 2 variables deviate in the same direction

i.e., if the increase in the values of one variable results in a corresponding increase in the values of other variable (or) if the decrease in the values of one variable results in a corresponding decrease in the values of other variable is called Positive Correlation.

e.g. Heights & weights of the individuals If the increase (decrease) in the values of one variable results in a corresponding decrease (increase) in the values of other variable is called Negative Correlation.

e.g, Price and demand of a commodity.

2) Linear and Non-linear Correlation: The correlation between two variables is said to be Linear if the corresponding to a unit change in one variable there is a constant change in the other variable over the entire range of the values (or) two variables x, y are said to be linearly related if there exists a relationship of the form $y = a + bx$.

e.g when the amount of output in a factory is doubled by doubling the number of workers.

Two variables are said to be Non linear or curvilinear if corresponding to a unit change

in one variable the other variable does not change at a constant rate but at fluctuating rate.

i.e Correlation is said to be non linear if the ratio of change is not constant. In other words, when all the points on the scatter diagram tend to lie near a smooth curve, the correlation is said   to   be   non   linear   (curvilinear).

3) Partial and Total correlation: The study of two variables excluding some other variables   is called Partial correlation .

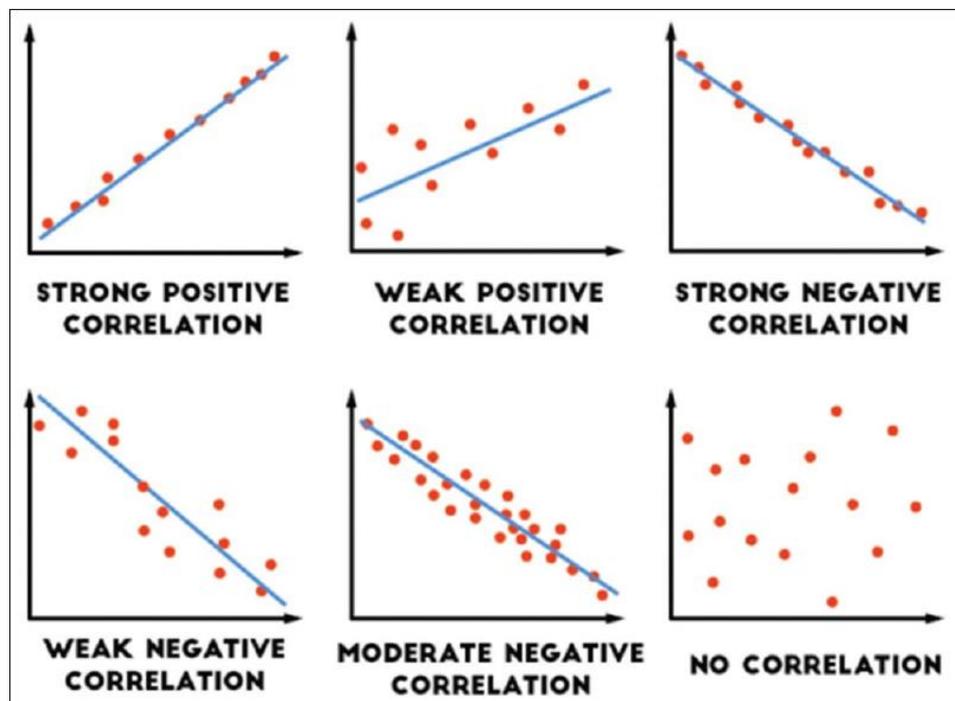e.g. We study price and demand eliminating the supply.

In Total correlation all the facts are taken into account.

e.g Price, demand & supply ,all are taken into account.

4) Simple and Multiple correlation: When we study only two variables, the relationship is described as Simple correlation.

E.g quantity of money and price level, demand and price.

The following are scatter diagrams of Correlation.

STRONG POSITIVE
CORRELATION

WEAK POSITIVE
CORRELATION

STRONG NEGATIVE
CORRELATION

WEAK NEGATIVE
CORRELATION

MODERATE NEGATIVE
CORRELATION

NO CORRELATION

Karl Pearson's Coefficient of Correlation

Karl Pearson suggested a mathematical method for measuring the magnitude of linear relationship between 2 variables. This is known as Pearsonian Coefficient of correlation. It is denoted by 'r'. This method is also known as Product-Moment correlation coefficient

$$r = \frac{Cov(xy)}{\sigma_x \sigma_y}$$

$$= \frac{\sum xy}{N\sigma_x \sigma_y}$$

$$= \frac{\sum XY}{\sqrt{\sum X^2 \sum Y^2}}$$

$X = (x - \overline{X})$, $Y = (y - \overline{Y})$ where, $\overline{X}$, $\overline{Y}$ are means of the series x & y.

$\sigma_x = $ standard deviation of series x

$\sigma_y = $ standard deviation of series y

Properties

1. The Coefficient of correlation lies b/w -1 & $+1$.
2. The Coefficient of correlation is independent of change of origin & scale of measurements.
3. If X, Y are random variables and a, b, c, d are any numbers such that $a \neq 0, c \neq 0$ then

$$r(aX + b, cY + d) = \frac{ac}{|ac|} r(X, Y)$$

4. Two independent variables are uncorrelated. That is if X and Y are independent variables then $r(X, Y) = 0$

Rank Correlation Coefficient

Charles Edward Spearman found out the method of finding the Coefficient of correlation by ranks. This method is based on rank & is useful in dealing with qualitative characteristics such as morality, character, intelligence and beauty. Rank correlation is applicable to only to the individual observations.

formula: $\rho = 6 \frac{\sum D^2}{N(N^2 - 1)}$

where : ρ - Rank Coefficient of correlation

D^2 - Sum of the squares of the differences of two ranks

N- Number of paired observations.

Properties

1. The value of ρ lies between $+1$ and -1.
2. If $\rho = 1$, then there is complete agreement in the order of the ranks & the direction of the rank is same.
3. If $\rho = -1$, then there is complete disagreement in the order of the ranks & they are in opposite directions.

Equal or Repeated ranks

If any 2 or more items are with same value the in that case common ranks are given to repeated items. The common rank is the average of the ranks which these items would have assumed, if they were different from each other and the next item will get the rank next to ranks already assumed.

Formula: $\rho = 1 - 6\{\dfrac{\sum D^2 + \frac{1}{12}(m^3 - m) + \frac{1}{12}(m^3 - m)....}{N^3 - N}\}$

where m $=$ the number of items whose ranks are common.

N- Number of paired observations.

D^2 - Sum of the squares of the differences of two ranks

REGRESSION

In regression we can estimate value of one variable with the value of the other variable which is known. The statistical method which helps us to estimate the unknown value of one variable from the known value of the related variable is called 'Regression'. The line described in the average relationship b/w 2 variables is known as <u>Line of Regression.</u>
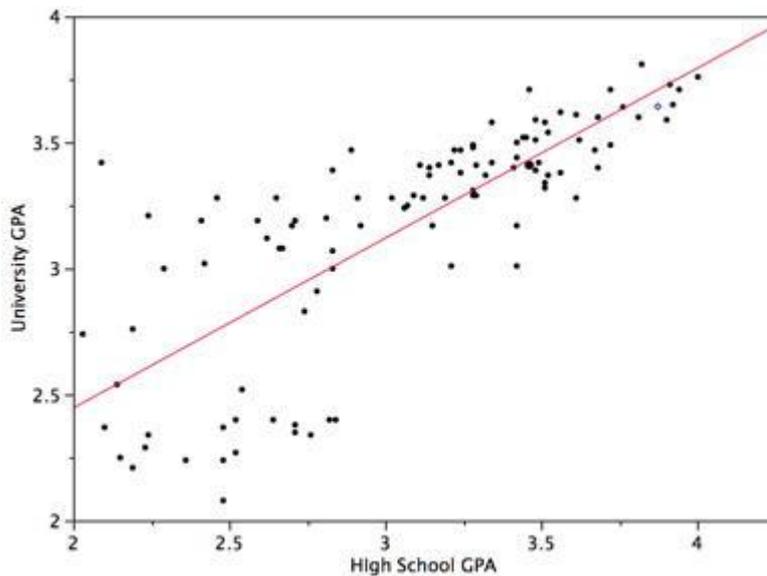
Regression Equation:

The standard form of the Regression equation is $Y = a + bX$ where a, b are called constants. 'a' indicates value of Y when $X = 0$. It is called Y-intercept. 'b' indicates the value of slope of the regression line & gives a measure of change of y for a unit change in X . it is also called as regression coefficient of Y on X. The values of a, b are found with the help of following <u>Normal Equations</u>.

Normal Equations for Regression Equation of Y on X are

$$\sum Y = Na + b\sum X$$

$$\sum XY = a \sum X + b\sum X^2$$

Normal Equations for Regression Equation of X on Y are

$$\sum X = Na + b\sum Y$$

$$\sum XY = a \sum Y + b\sum Y^2$$

Regression equations when deviations taken from the arithmetic mean :

Regression equation of Y on X : $Y - \overline{Y} = b_{yx}(X - \overline{X})$ where $b_{yx} = \frac{\sum XY}{\sum X^2}$

Regression equation of X on Y : $X - \overline{X} = b_{xy}(Y - \overline{Y})$ where $b_{xy} = \frac{\sum XY}{\sum Y^2}$

Angle b/w Two Regression lines : $\tan\theta = \frac{m_1 - m_2}{1 + m_1 m_2}$

Note:

1. If θ is acute then $\tan\theta = \frac{\sigma_x \sigma_y}{\sigma^2_x + \sigma^2_y}\left(\frac{1 - r^2}{r}\right)$

2. If θ is obtuse then $\tan\theta = \frac{\sigma_x \sigma_y}{\sigma^2_x + \sigma^2_y}\left(\frac{r^2 - 1}{r}\right)$

3. If $r = 0$ then $\tan\theta = \infty$ then $\theta = \frac{\pi}{2}$. Thus if there is no relationship between the 2 variables (i.e, they are independent) then $\theta = \frac{\pi}{2}$.

4. If $r = \pm 1$ then $\tan\theta = 0$ then $\theta = 0$ or π. Hence the 2 regression lines are parallel or coincident. The correlation between 2 variables is perfect.

Problems

1. **Find Karl Pearson's coefficient of correlation from the following data.**

Ht. in inches	57	59	62	63	64	65	55	58	57
Weight in lbs	113	117	126	126	130	129	111	116	112

Solution:

Ht. in inches X	Deviation from mean $X = x - \bar{x}$	X^2	Wt. in lbs Y	Deviation from mean $Y = y - \bar{y}$	Y^2	Product of deviations of X and Y series (XY)
57	-3	9	113	-7	49	21
59	-1	1	117	-3	9	3
62	2	4	126	6	36	12
63	3	9	126	6	36	18
64	4	16	130	10	100	40
65	5	25	129	9	81	45
55	-5	25	111	-9	81	45
58	-2	4	116	-4	16	8
57	-3	9	112	-8	64	24
540	0	102	1080	0	472	216

Coefficient of correlation $r = \dfrac{\sum XY}{\sqrt{\sum X^2 \sum Y^2}} = \dfrac{216}{\sqrt{(102)(471)}} = 0.98$

2. **Calculate Coefficient of correlation for the following data.**

X	12	9	8	10	11	13	7
Y	14	8	6	9	11	12	3

Solution: In both series items are in small number.

So there is no need to take deviations.

$$\text{Formula used: } r = \frac{\text{Cov(XY)}}{\sigma_x \sigma_y}$$

X	Y	X²	Y²	XY
12	14	144	196	168
9	8	81	64	72
8	6	64	36	48
10	9	100	81	90
11	11	121	121	121
13	12	169	144	156
7	3	49	9	21
$\sum X = 70$	$\sum Y = 63$	$\sum X^2 = 728$	$\sum Y^2 = 651$	$\sum XY = 676$

$$r = \frac{\sum XY - (\sum X \sum Y)/N}{\sqrt{(\sum X^2) - (\sum X)^2/N)(\sum Y^2 - (\sum Y)^2)/N}}$$

Here $N = 7$.

$$r = \frac{4732 - 4410}{\sqrt{5096 - 4900}\sqrt{4557 - 3969}} = \frac{322}{\sqrt{(196)(588)}} = \frac{322}{339.48} = +0.95$$

3. **A sample of 12 fathers and their elder sons gave the following data about their elder sons. Calculate the rank correlation coefficient.**

Fathers	65	63	67	64	68	62	70	66	68	67	69	71
Sons	68	66	68	65	69	66	68	65	71	67	68	70

Solution:

Fathers(x)	Sons(y)	Rank(x)	Rank(y)	$d_i = x_i - y_i$	d_i^2
65	68	9	5.5	3.5	12.25
63	66	11	9.5	1.5	2.25
67	68	6.5	5.5	1.0	1

64	65	10	11.5	-1.5	2.25
68	69	4.5	3	1.5	2.25
62	66	12	9.5	2.5	6.25
70	68	2	5.5	=3.5	12.25
66	65	8	11.5	3.5	12.25
68	71	4.5	1	=3.5	12.25
67	67	6.5	8	-1.5	2.25
69	68	3	5.5	-2.5	6.25
71	70	1	2	-1	1
					$\sum d_i^2 = 72.5$

Repeated values are given common rank, which is the mean of the ranks .In X: 68 & 67 appear twice.

In Y : 68 appears 4 times , 66 appears twice & 65 appears twice. Here N $= 12$.

$$\rho = 1 - 6\left\{\frac{\sum D^2 + \frac{1}{12}(m^3 - m) + \frac{1}{12}(m^3 - m)}{N^3 - N}\right\} = 1 - \frac{6(72.5 + 7)}{12(12^2 - 1)} = 0.722$$

4. **Given n $= 10$, $\sigma_x = 5.4$, $\sigma_y = 6.2$ and sum of product of deviation from the mean of X & Y is 66. Find the correlation coefficient.**

Solution: n $= 10$, $\sigma_x = 5.4, \sigma_y = 6.2$

$$\sigma_x^2 = \frac{\sum(x - \bar{x})^2}{n}$$

$$\sigma_y^2 = \frac{\sum(y - \bar{y})^2}{n}$$

$$\sum(x - \bar{x})(y - \bar{y}) = 66$$

$$r = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sqrt{\sum(x - \bar{x})^2 \sum(y - \bar{y})^2}} = \frac{66}{(5.)(6.2)} = 0.1971$$

5. The heights of mothers & daughters are given in the following table. From the 2 tables of   regression estimate the expected average height of daughter when the height of the mother   is 64. 5 inches.

Ht. of Mother(inches)	62	63	64	64	65	66	68	70

Ht. of the daughter(inches)	64	65	61	69	67	68	71	65

Solution:

Let X = heights of the mother

Y = heights of the daughter

Let $dx = X - 65$, $dy = Y - 67$, $\sum x = 522$, $\sum dx = 2$, $\sum dx^2 = 50$, $\sum y = 530$,

$$\sum dy = -6 \quad \sum dy^2 = 74, \quad \sum dxdy = 20$$

$$\bar{X} = \frac{\sum X}{N} = \frac{522}{8} = 66.25$$

$$\bar{Y} = \frac{\sum Y}{N} = \frac{530}{8} = 65.25$$

$$b_{yx} = \frac{\frac{\sum dxdy - \sum dx \sum dy}{N}}{\sum dx^2 - \frac{(\sum dx)^2}{N}} = \frac{20 - \frac{2(-6)}{8}}{50 - \frac{2}{8}} = 0.434$$

Regression equation of Y on : $Y - \bar{Y} = b_{yx}(X - \bar{X})$

$$Y = 37.93 + 0.434X$$

when $X = 64.5$ then $Y = 69.923$

6. The equations of two regression lines are $7x - 16y + 9 = 0$ and $5y - 4x - 3 = 0$.

Find the coefficient of correlation and the means of x& y.

Solution: Given equations are $7x - 16y + 9 = 0$……………….. (1)

$$5y - 4x - 3 = 0………………….(2)$$

(1) × 4 gives $28x - 64y + 36 = 0$

(2) × 7 gives $-28x + 35y - 21 = 0$

On adding we get $-29y + 15 = 0$

$y = 0.5172$

from(1) $7x = 16y - 9$ which gives $x = 0.1034$

since regression line passes through $(\bar{x}, \bar{y})$ we have $\bar{x} = 0.1034$

$\bar{y} = 0.5172$

From(1) $x = \frac{16}{7}y - \frac{9}{7}$

From (2) $y = \frac{4}{5}x + \frac{3}{5}$,

$r\frac{\sigma_x}{\sigma_y} = \frac{16}{7}$ and $r\frac{\sigma_y}{\sigma_x} = \frac{4}{5}$

Multiplying these 2 equations , we get $r^2 = \frac{16}{7}\frac{4}{5} = \frac{64}{35}$

$r = \frac{8}{\sqrt{35}}$.

7. If $\sigma_x = \sigma_y = \sigma$ and the angle between the regression lines is $Tan^{-1}\left(\frac{4}{3}\right)$. Find r.

Solution: $\tan\theta = \frac{\sigma_x\sigma_y}{\sigma^2_x + \sigma^2_y}\left(\frac{1-r^2}{r}\right)$

$$= \frac{\sigma^2}{2\sigma^2}\left(\frac{1-r^2}{r}\right)$$

By data, $\theta = Tan^{-1}\left(\frac{4}{3}\right)$.

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

$3 - 3r^2 - 8r = 0$

$(3r - 1)(r + 3) = 0$

$r = \frac{1}{3}$ or -3

Since we cannot have $r = -3$

Thus $r = \frac{1}{3}$

8. Given the following information regarding a distribution $N = 5$,

$\overline{X} = 10, \overline{Y} = 20, \sum(X - Y)^2 = 100, \sum(Y - 10)^2 = 160$. **Find the regression**

coefficients and hence coefficient of correlation.

Solution: Here $dx = X - 4, dy = Y - 10$

$$\overline{X} = A + \frac{\sum dx}{N} \Rightarrow 10 = Y + \frac{\sum dx}{5} \Rightarrow \sum dx = 30 (\text{here } A = 4)$$

$$\bar{Y} = B + \frac{\sum dy}{N} \Rightarrow 20 = 10 + \frac{\sum dy}{5} \Rightarrow \sum dy = 50 \text{(here } B = 10)$$

$$b_{yx} = \frac{\sum dxdy - \frac{\sum dx \sum dy}{N}}{\sum dx^2 - \frac{(\sum dx)^2}{N}} = \frac{-220}{-80} = 2.75$$

$$b_{xy} = \frac{\sum dxdy - \frac{\sum dx \sum dy}{N}}{\sum dy^2 - \frac{(\sum dy)^2}{N}} = \frac{-220}{-340} = 0.65$$

Coefficient of correlation $r = \pm\sqrt{b_{xy} \times b_{yx}} = \sqrt{(0.65)(2.75)} = \sqrt{1.7875} = 1.337$

9. Given that X = 4Y + 5 and Y = 4X + 4 are the lines of regression of X on Y and Y on X respectively. Show that $0 < 4k < 1$. If $k = \frac{1}{16}$ find the means of the two variables and coefficient of correlation between them.

Solution: Given lines are X= 4Y +5 …….(1)

$$Y = KX+4 \text{ ……….(2)}$$

From (1) & (2), $r\frac{\sigma_x}{\sigma_y} = 4$ and $r\frac{\sigma_y}{\sigma_x} = K$

Multiplying these two equations we get $r^2 = 4K$

Since $0 \leq r^2 \leq 1$, we have $0 \leq 4K \leq \frac{1}{4}$

If $K = \frac{1}{16}$ then we have $X = 4Y + 5$ and

$$Y = X/16 + 4$$

We get $X - 4Y - 5 = 0$

$$\frac{-X}{4}4Y - 16 = 0$$

Adding we get $3\frac{X}{4} - 21 = 0$

$X = 28$

From(2), we get $Y = \frac{23}{4}$

The regression lines pass through $(\bar{x}, \bar{y})$

We get means $\bar{x} = 28$ and $\bar{y} = \frac{23}{4}$

We have $r^2 = 4k = \frac{4}{16} = \frac{1}{4} \Rightarrow r = \pm \frac{1}{2}$

We consider positive value and take $r = \frac{1}{2}$

10. The difference between the ranks are $0.5, -6, -4.5, -3, -5, -1, 3, 0, 5, 5.5, 0, -0.5$.

For refracted ranks x and y. $\frac{\sum m(m^2-1)}{12}=3.5, r = 0.44$. Find the number of terms.

Solution: Given difference (d_i) $0.5, -6, -4.5, -3, -5, -1, 3, 0, 5, 5.5, 0, -0.5$

$$\sum d_i^2 = 156$$

Here $r = 1 - 6\left\{ \dfrac{\sum d_i^2 + \frac{\sum m(m^2-1)}{12}}{(N^2-N)} \right\}$

$$= \frac{1-(159.5)6}{(N^2-N)}$$

$$= 1 - \frac{957}{N^2-N}$$

$$\Rightarrow 0.44 = 1 - \frac{957}{N^2-N}$$

$$\Rightarrow N^2 - N = 1708.92$$

$$\Rightarrow N = 42$$

1. The heights of mothers & daughters are given in the following table. From the 2 tables of regression estimate the expected average height of daughter when the height of the mother is 64.5 inches.

Ht. of Mother(inches)	62	63	64	64	65	66	68	70
Ht. of the daughter(inches)	64	65	61	69	67	68	71	65

2. The equations of two regression lines are $7x - 16y + 9 = 0$ and $5y - 4x - 3 = 0$.

   Find the coefficient of correlation and the means of x & y .

3. The marks obtained by 10 students in mathematics and statistics are given below. Find the coefficient of correlation between the two subjects and the two lines of regression

Marks in mathematics	25	28	30	32	35	36	38	42	45	39
Marks in Statistics	20	26	29	30	25	18	26	35	46	35

4. Fit a straight line $Y=a_0+a_1X$ for the following data and estimate the value of Y when X=25

X	0	5	10	15	20
Y	7	11	16	20	26

5. Find the rank correlation for the following indices of supply and price of an article:

PRICE	80	100	102	91	100	111	109	100	99	104	111	102	98	111
INDEX	124	100	105	112	102	93	99	115	123	104	99	113	121	103

ASSIGNMENT QUESTIONS

1.Fit a curve of the form Y= a +bX by the method of least squares for the following data:

X	1	2	3	4	5
Y	5	2	4.5	8	12.5

\
2.The marks obtained by 10 students in two subjects are given below. Find the correlation coefficient and lines of regression

Subject 1	48	75	30	60	80	53	35	15	40	38
Subject 2	44	85	45	54	91	58	63	35	43	45

3. The following table are the marks obtained by 12 students in economics and statistics:

Economics(X)	78	56	36	66	25	62	75	82
62Statistics(Y)	84	44	51	58	60	58	68	62

Obtain the regression lines.

4.Find the karlpearson's coefficient of correlation for the paired data:

wages	100	101	102	100	99	97	98	96	95	102
Cost of living	98	99	99	95	92	95	94	90	91	97

5. The equations of two regression lines are $7x - 16y + 9 = 0$ and $5y - 4x - 3 = 0$.

   Find the coefficient of correlation and the means of x& y .

UNIT 4
SAMPLING

OBJECTIVE

Know about the sampling distribution of a given statistic of a

 random sample.

OUTCOME

Understand the importance of sampling distribution of a given statistic of a

 random sample

SAMPLING

Introduction: The totality of observations with which we are concerned , whether this number be finite or infinite constitute population. In this chapter we focus on sampling from distributions or populations and such important quantities as the sample mean and sample variance.

Def: Population is defined as the aggregate or totality of statistical data forming a subject of investigation .

EX. The population of the heights of Indian.

The number of observations in the population is defined to be the size of the population. It may be finite or infinite .Size of the population is denoted by N.As the study of entire population may not be possible to carry out and hence a part of the population alone is selected.

Def: A portion of the population which is examined with a view to determining the population characteristics is called a sample . In other words, sample is a subset of population. Size of the sample is denoted by n.

The process of selection of a sample is called Sampling. There are different methods of sampling

- ➢ Probability Sampling Methods
- ➢ Non-Probability Sampling Methods

Probability Sampling Methods:

a) **Random Sampling (Probability Sampling**):
   It is the process of drawing a sample from a population in such a way that each member of the population has an equal chance of being included in the sample.
   Ex: A hand of cards from a well shuffled pack of cards is a random sample.
   Note : If N is the size of the population and n is the size of the sample, then
      - ➢ The no. of samples with replacement $= N^n$
      - ➢ The no. of samples without replacement $= N_{C_n}$

b) **Stratified Sampling :**
   In this , the population is first divided into several smaller groups called strata according to some relevant characteristics . From each strata samples are selected at random, all the samples are combined together to form the stratified sampling.

c) **Systematic Sampling (Quasi Random Sampling):**
   In this method , all the units of the population are arranged in some order . If the population size is N, and the sample size is n, then we first define sample interval denoted by $= \frac{N}{n}$. then from first k items ,one unit is selected at random. Then from first unit every k^{th} unit is serially selected combining all the selected units constitute a systematic sampling.

Non Probability Sampling Methods:

a) **Purposive (Judgment) Sampling :**
In this method, the members constituting the sample are chosen not according to some definite scientific procedure , but according to convenience and personal choice of the individual who selects the sample . It is the choice of the individual items of a sample entirely depends on the individual judgment of the investigator.

b) **Sequential Sampling:**
It consists of a sequence of sample drawn one after another from the population. Depending on the results of previous samples if the result of the first sample is not acceptable then second sample is drawn and the process continues to take proper decision . But if the first sample is acceptable ,then no new sample is drawn.

Classification of Samples:

➢ **Large Samples :** If the size of the sample $n \geq 30$, then it is said to be large sample.

➢ **Small Samples :** If the size of the sample $n < 30$,then it is said to be small sample or exact sample.

Parameters and Statistics:

Parameter is a statistical measure based on all the units of a population. Statistic is a statistical measure based on only the units selected in a sample.

Note :In this unit , Parameter refers to the population and Statistic refers to sample.

Central Limit Theorem: If $\bar{x}$ be the mean of a random sample of size n drawn from population having mean μ and standard deviation σ , then the sampling distribution of the sample mean $\bar{x}$ is approximately a normal distribution with mean μ and SD = S.E of $\bar{x} = \frac{\sigma}{\sqrt{n}}$ provided the sample size n is large.

Standard Error of a Statistic : The standard error of statistic 't' is the standard deviation of the sampling distribution of the statistic i.e, S.E of sample mean is the standard deviation of the sampling distribution of sample mean.

Formulae for S.E:

➢ S.E of Sample mean $\bar{x} = \frac{\sigma}{\sqrt{n}}$ i.e, S.E $(\bar{x}) = \frac{\sigma}{\sqrt{n}}$

➢ S.E of sample proportion p$= \sqrt{\frac{PQ}{n}}$ i.e, S.E (p) $= \sqrt{\frac{PQ}{n}}$ where Q=1-P

➢ S.E of the difference of two sample means $\overline{x_1}$ and $\overline{x_2}$ i.e, S.E $(\overline{x_1} - \overline{x_2}) = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

➢ S.E of the difference of two proportions i.e, S.E$(p_1 - p_2) = \sqrt{\frac{P_1 Q_1}{n_1} + \frac{P_2 Q_2}{n_2}}$

Estimation :

To use the statistic obtained by the samples as an estimate to predict the unknown parameter of the population from which the sample is drawn.

Estimate : An estimate is a statement made to find an unknown population parameter.

Estimator : The procedure or rule to determine an unknown population parameter is called estimator.

Ex. Sample proportion is an estimate of population proportion , because with the help of sample proportion value we can estimate the population proportion value.

Types of Estimation:

➢ **Point Estimation:** If the estimate of the population parameter is given by a single value , then the estimate is called a point estimation of the parameter.

➢ **Interval Estimation:** If the estimate of the population parameter is given by two different values between which the parameter may be considered to lie, then the estimate is called an interval estimation of the parameter.

Confidence interval Estimation of parameters:

In an interval estimation of the population parameter θ, if we can find two quantities t_1 and t_2 based on sample observations drawn from the population such that the unknown parameter θ is included in the interval $[t_1, t_2]$ in a specified cases ,then this is called a confidence interval for the parameter θ.

Confidence Limits for Population mean μ

➢ 95% confidence limits are $\bar{x} \pm 1.96\ (S.E.\ of\ \bar{x})$
➢ 99% confidence limits are $\bar{x} \pm 2.58\ (S.E.\ of\ \bar{x})$
➢ 99.73% confidence limits are $\bar{x} \pm 3\ (S.E.\ of\ \bar{x})$
➢ 90% confidence limits are $\bar{x} \pm 1.645\ (S.E.\ of\ \bar{x})$

Confidence limits for population proportion P

➢ 95% confidence limits are $p \pm 1.96(S.E.of\ p)$
➢ 99% confidence limits are $p \pm 2.58(S.E.\ of\ p)$
➢ 99.73% confidence limits are $p \pm 3(S.E.of\ p)$
➢ 90% confidence limits are $p \pm 1.645(S.E.of\ p)$

Confidence limits for the difference of two population means μ_1 and μ_2

➢ 95% confidence limits are $((\overline{x_1} - \overline{x_2}) \pm 1.96\ (S.E\ of\ ((\overline{x_1} - \overline{x_2})))$
➢ 99% confidence limits are $((\overline{x_1} - \overline{x_2}) \pm 2.58\ (S.E\ of\ ((\overline{x_1} - \overline{x_2})))$
➢ 99.73% confidence limits are $((\overline{x_1} - \overline{x_2}) \pm 3\ (S.E\ of\ ((\overline{x_1} - \overline{x_2})))$
➢ 90% confidence limits are $((\overline{x_1} - \overline{x_2}) \pm 2.58\ (S.E\ of\ ((\overline{x_1} - \overline{x_2})))$

Confidence limits for the difference of two population proportions

➢ 95% confidence limits are p_1-p_2 $\pm$ 1.96 (S.E. of p_1-p_2)
➢ 99% confidence limits are p_1-p_2 $\pm$ 2.58 (S.E. of p_1-p_2)
➢ 99.73% confidence limits are p_1-p_2 $\pm$ 3 (S.E. of p_1-p_2)
➢ 90% confidence limits are p_1-p_2 $\pm$ 1.645 (S.E. of p_1-p_2)

Determination of proper sample size

Sample size for estimating population mean :

$n = \left(\frac{z_\alpha \sigma}{E}\right)^2$ where z_α – Critical value of z at α Level of significance

σ – Standard deviation of population and

E – Maximum sampling Error $= \bar{x} - \mu$

Sample size for estimating population proportion :

$n = \frac{z_\alpha{}^2 PQ}{E^2}$ where z_α – Critical value of z at α Level of significance

P – Population proportion
Q – 1-P
E – Maximum Sampling error = p-P

Testing of Hypothesis :

It is an assumption or supposition and the decision making procedure about the assumption whether to accept or reject is called hypothesis testing .

Def: Statistical Hypothesis : To arrive at decision about the population on the basis of sample information we make assumptions about the population parameters involved such assumption is called a statistical hypothesis .

Procedure for testing a hypothesis:

Test of Hypothesis involves the following steps:

Step1: Statement of hypothesis :

There are two types of hypothesis :

➢ **Null hypothesis:** A definite statement about the population parameter. Usually a null hypothesis is written as no difference , denoted by H_0.
Ex. $H_0: \mu = \mu_0$
➢ **Alternative hypothesis :** A statement which contradicts the null hypothesis is called alternative hypothesis. Usually an alternative hypothesis is written as some difference , denoted by H_1.

Setting of alternative hypothesis is very important to decide whether it is two-tailed or one – tailed alternative , which depends upon the question it is dealing.

Ex. $H_1 : \mu \neq \mu_0$ (Two – Tailed test)

 or

$H_1 : \mu > \mu_0$ (Right one tailed test)

 or

$H_1 : \mu < \mu_0$ (Left one tailed test)

Step 2: Specification of level of significance :

The LOS denoted by α is the confidence with which we reject or accept the null hypothesis. It is generally specified before a test procedure ,which can be either 5% (0.05) , 1% or 10% which means that thee are about 5 chances in 100 that we would reject the null hypothesis H_0 and the remaining 95% confident that we would accept the null hypothesis H_0 . Similarly , it is applicable for different level of significance.

Step 3 : Identification of the test Statistic :

There are several tests of significance like z,t, F etc .Depending upon the nature of the information given in the problem we have to select the right test and construct the test criterion and appropriate probability distribution.

Step 4: Critical Region:

It is the distribution of the statistic .

➢ **Two – Tailed Test : The critical region under the curve is equally distributed on both sides of the mean.**

If H_1 has $\neq$ sign , the critical region is divided equally on both sides of the distribution.
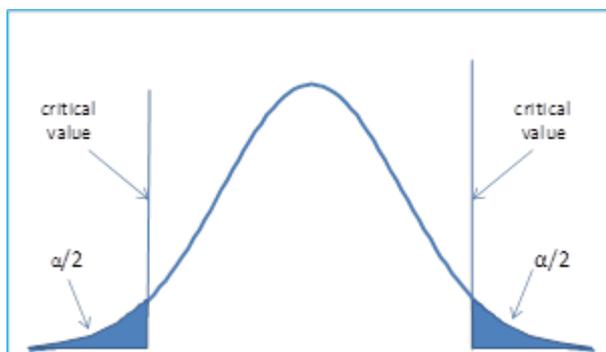
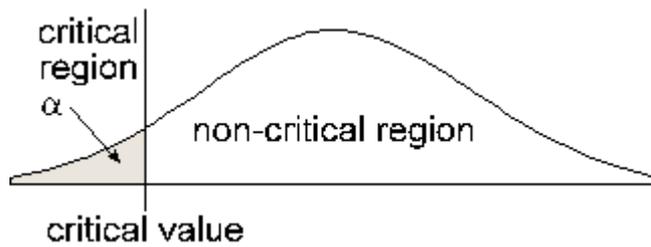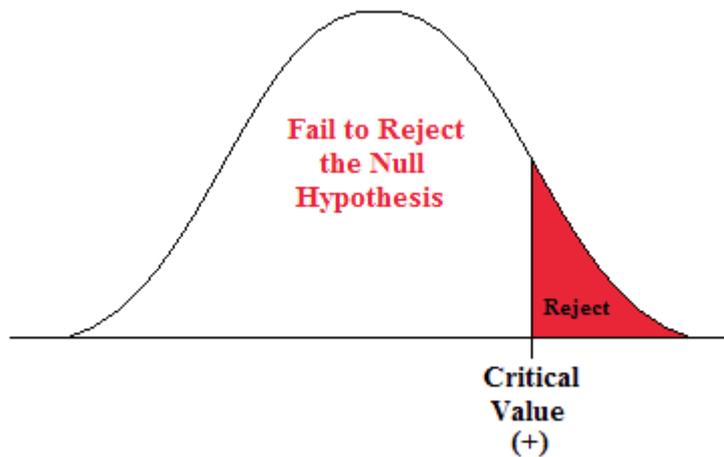➢ **One Tailed Test: The critical region under the curve is distributed on one side of the mean.**

Left one tailed test: If H_1 has $<$ sign , the critical region is taken in the left side of the distribution.

Right one tailed test : If H_1 has $>$ sign , the critical region is taken on right side of the distribution.

Step 5 : Making decision:

By comparing the computed value and the critical value decision is taken for accepting or rejecting H_0

If calculated value $\leq$ critical value , we accept H_0, otherwise reject H_0.

Errors of Sampling :

While drawing conclusions for population parameters on the basis of the sample results , we have two types of errors.

➢ **Type I error :** Reject H_0 when it is true i.e, if the null hypothesis H_0 is true but it is rejected by test procedure .
➢ **Type II error :** Accept H_0 when it is false i.e, if the null hypothesis H_0 is false but it is accepted by test procedure.

DECISION TABLE

	H_0 is accepted	H_0 is rejected
H_0 is true	Correct Decision	Type I Error
H_0 is false	Type II Error	Correct Decision

Problems:

1. **If the population is 3,6,9,15,27**

    a) **List all possible samples of size 3 that can be taken without replacement from finite population**

    b) **Calculate the mean of each of the sampling distribution of means**

    c) **Find the standard deviation of sampling distribution of means**

Sol: Mean of the population , $\mu = \frac{3+6+9+15+27}{5} = \frac{60}{5} = 12$

Standard deviation of the population ,

$$\sigma = \sqrt{\frac{(3-12)^2 + (6-12)^2 + (9-12)^2 + (15-12)^2 + (27-12)^2}{5}}$$

$$= \sqrt{\frac{81+36+9+9+225}{5}} = \sqrt{\frac{360}{5}} = 8.4853$$

a) Sampling without replacement :

The total number of samples without replacement is $N_{C_n} = 5_{C_3} = 10$

The 10 samples are (3,6,9), (3,6,15), (3,9,15), (3,6,27), (3,9,27), (3,15,27), (6,9,15), (6,9,27), (6,15,27), (9,15,27)

b) Mean of the sampling distribution of means is

$$\mu_{\bar{x}} = \frac{6+8+9+10+12+13+14+15+16+17}{10} = \frac{120}{10} = 12$$

c) $\sigma^2 = $

$$\frac{(6-12)^2 + (8-12)^2 + (9-12)^2 + (10-12)^2 + (12-12)^2 + (13-12)^2 + (14-12)^2 + (15-12)^2 + (16-12)^2 + (17-12)^2}{10}$$

$$= 13.3$$

$$\therefore \sigma_{\bar{x}} = \sqrt{13.3} = 3.651$$

2. **A population consist 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 standard deviation of the population**

    c) **The mean of the sampling distribution of means and**

    d) **The standard deviation of the sampling distribution of means**

Sol: a) Mean of the Population is given by

$$\mu = \frac{2+3+6+8+11}{5} = \frac{30}{5} = 6$$

b) Variance of the population is given by

$$\sigma^2 = \sum \frac{(x_i - \bar{x})^2}{n}$$

$$= \frac{(2-6)^2+(3-6)^2+(6-6)^2+(8-6)^2+(11-6)^2}{5}$$

$$= \frac{16+9+0+4+25}{5} = 10.8 \qquad \therefore \sigma = 3.29$$

c) Sampling with replacement

The total no.of samples with replacement is $N^n = 5^2 = 25$

$\therefore$ List of all possible samples with replacement are

$$\begin{Bmatrix} (2,2),(2,3),(2,6),(2,8),(2,11),(3,2),(3,3)(3,6),(3,8),(3,11) \\ (6,2),(6,3),(6,6),(6,8),(6,11),(8,2),(8,3),(8,6),(8,8),(8,11) \\ (11,2),(11,3),(11,6),(11,8),(11,11) \end{Bmatrix}$$

Now compute the arithmetic mean for each of these 25 samples which gives rise to the distribution of means of the samples known as sampling distribution of means

The samples means are

$$\begin{Bmatrix} 2,2.5,4,5,6.5 \\ 2.5,3,4.5,,5.5,7 \\ 4,4.5,6,7,8.5 \\ 5,5.5,7,8,9.5 \\ 6.5,7,8.5,9.5,11 \end{Bmatrix}$$

And the mean of sampling distribution of means is the mean of these 25 means

$$\mu_{\bar{x}} = \frac{\text{sum of all above sample means}}{25} = \frac{150}{25} = 6$$

d) The variance of the sampling distribution of means is obtained by subtracting the mean 6 from each number in sampling distribution of means and squaring the result ,adding all 25 numbers thus obtained and dividing by 25.

$$\sigma^2 = \frac{(2-6)^2+(2.5-6)^2+(4-6)^2+(5-6)^2+\cdots\ldots(11-6)^2}{25} = \frac{135}{25} = 5.4$$

$$\therefore \sigma = \sqrt{5.4} = 2.32$$

3.When a sample is taken from an infinite population , what happens to the standard error of the mean if the sample size is decreased from 800 to 200

Sol: The standard error of mean $= \frac{\sigma}{\sqrt{n}}$

Sample size $= n$ .let n$= n_1 = 800$

Then S.E$_1 = \frac{\sigma}{\sqrt{800}} = \frac{\sigma}{20\sqrt{2}}$

When n_1 is reduced to 200

let n$= n_2 = 200$

Then S.E$_2 = \frac{\sigma}{\sqrt{200}} = \frac{\sigma}{10\sqrt{2}}$

$$\therefore S.E_2 = \frac{\sigma}{10\sqrt{2}} = 2\left(\frac{\sigma}{20\sqrt{2}}\right) = 2\ (S.E_1)$$

Hence if sample size is reduced from 800 to 200, S. E. of mean will be multiplied by 2

4. The variance of a population is 2 . The size of the sample collected from the population is 169. What is the standard error of mean

Sol: n= The size of the sample =169

$$\sigma = \text{S.D of population} = \sqrt{\text{Variance}} = \sqrt{2}$$

$$\text{Standard Error of mean} = \frac{\sigma}{\sqrt{n}} = \frac{\sqrt{2}}{\sqrt{169}} = \frac{1.41}{13} = 0.185$$

5. The mean height of students in a college is 155cms and standard deviation is 15 . What is the probability that the mean height of 36 students is less than 157 cms.

Sol: μ = Mean of the population

= Mean height of students of a college = 155cms

n = S.D of population = 15cms

$\bar{x}$ = mean of sample = 157 cms

$$\text{Now } z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{157 - 155}{\frac{15}{\sqrt{36}}} = \frac{12}{15} = 0.8$$

$$\therefore P\ (\ \bar{x} \leq 157) = P\ (\ z < 0.8\ ) = 0.5 + P\ (\ 0 \leq z \leq 0.8\ )$$

$$= 0.5 + 0.2881 = 0.7881$$

Thus the probability that the mean height of 36 students is less than 157 = 0.7881

6. A random sample of size 100 is taken from a population with $\sigma = 5.1$. Given that the sample mean is $\bar{x} = 21.6$ Construct a 95% confidence limits for the population mean .

Sol: Given $\bar{x} = 21.6$

$z\alpha_{/2} = 1.96$, n = 100 , $\sigma = 5.1$

$\therefore$ Confidence interval = $(\ \bar{x} - z\alpha_{/2}.\frac{\sigma}{\sqrt{n}}\ , \bar{x} + z\alpha_{/2}.\frac{\sigma}{\sqrt{n}}\ )$

$\bar{x} - z\alpha_{/2}.\frac{\sigma}{\sqrt{n}} = 21.6 - \frac{1.96 \text{ x } 5.1}{10} = 20.6$

$\bar{x} + z\alpha_{/2}.\frac{\sigma}{\sqrt{n}} = 21.6 + \frac{1.96 \text{ x } 5.1}{10} = 22.6$

Hence (20.6, 22.6) is the confidence interval for the population mean μ

7. It is desired to estimate the mean time of continuous use until an answering machine will first require service . If it can be assumed that $\sigma = 60$ days, how large a sample is

needed so that one will be able to assert with 90% confidence that the sample mean is off by at most 10 days.

Sol: We have maximum error (E) = 10 days , σ = 60 days and $z\alpha_{/2}$ = 1.645

$$\therefore n = \left[\frac{z\alpha_{/2} \cdot \sigma}{E}\right]^2 = \left[\frac{1.645 \times 60}{10}\right]^2 = 97$$

8.A random sample of size 64 is taken from a normal population with $\mu = 51.4$ and $\sigma = 6.8$. What is the probability that the mean of the sample will a) exceed 52.9   b) fall between 50.5 and 52.3 c) be less than 50.6

Sol: Given n = the size of the sample = 64

μ = the mean of the population = 51.4

σ = the S.D of the population = 6.8

a) P($\bar{x}$ exceed 52.9) = P($\bar{x}$ > 52.9)

$$Z = \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} = \frac{52.9-51.4}{\frac{6.8}{\sqrt{64}}} = 1.76$$

$\therefore$ P($\bar{x}$ > 52.9) = P(z > 1.76)

$= 0.5 - P(0 < z < 1.76)$
$= 0.5 - 0.4608 = 0.0392$

b) P($\bar{x}$ fall between 50.5 and 52.3)
   i.e, P(50.5 < $\bar{x}$ < 52.3) = P($\overline{x_1}$ < $\bar{x}$ < $\overline{x_2}$)
   $$z_1 = \frac{\overline{x_1}-\mu}{\frac{\sigma}{\sqrt{n}}} = \frac{50.5-51.4}{0.85} = -1.06$$

   $$z_2 = \frac{\overline{x_2}-\mu}{\frac{\sigma}{\sqrt{n}}} = \frac{52.3-51.4}{0.85} = 1.06$$
   P(50.5 < $\bar{x}$ < 52.3) = P(-1.06 < z < 1.06)
   = P(-1.06 < z < 0) + P(0 < z < 1.06)
   = P(0 < z < 1.06) + P(0 < z < 1.06)
   = 2(0.3554) = 0.7108

c) P($\bar{x}$ will be less than 50.6) = P($\bar{x}$ < 50.6)
   $$Z = \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} = \frac{50.6-51.4}{\frac{6.8}{\sqrt{64}}} = -0.94$$
   $\therefore$P(z < -0.94) = 0.5 - P(0.94 < z < 0)
   = 0.5 - P(0 < z < 0.94) = 0.50-0.3264
   = 0.1736

9.The mean of certain normal population is equal to the standard error of the mean of the samples of 64 from that distribution . Find the probability that the mean of the sample size 36 will be negative.

Sol: The Standard error of mean $= \dfrac{\sigma}{\sqrt{n}}$

Sample size , n =64

Given mean , $\mu =$ Standard error of the mean of the samples

$\mu = \dfrac{\sigma}{\sqrt{64}} = \dfrac{\sigma}{8}$

We know $z = \dfrac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \dfrac{\bar{x} - \frac{\sigma}{8}}{\frac{\sigma}{6}}$

$= \dfrac{6\bar{x}}{\sigma} - \dfrac{3}{4}$

If $Z < 0.75, \overline{\overline{x}}$ is negative

$P(z < 0.75) = P(-\infty < z < 0.75)$

$\qquad\qquad = \int_{-\infty}^{0} \emptyset(z)\, dz + \int_{0}^{0.75} \emptyset(z) dz = 0.50 + 0.2734$

$\qquad\qquad = 0.7734$

10.The guaranteed average life of a certain type of electric bulbs is 1500hrs with a S.D of 10 hrs. It is decided to sample the output so as to ensure that 95% of bulbs do not fall short of the guaranteed average by more than 2% . What will be the minimum sample size ?

Sol : Let n be the size of the sample

The guaranteed mean is 1500

We do not want the mean of the sample to be less than 2% of (1500) i.e, 30 hrs

So 1500 – 30 = 1470

$\therefore \bar{x} > 1470$

$\therefore |z| = \left| \dfrac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \right| = \left| \dfrac{1470 - 1500}{\frac{120}{\sqrt{n}}} \right| = \dfrac{\sqrt{n}}{4}$

From the given condition , the area of the probability normal curve to the left of $\dfrac{\sqrt{n}}{4}$ should be 0.95

$\therefore$ The area between 0 and $\dfrac{\sqrt{n}}{4}$ is 0.45

We do not want to know about the bulbs which have life above the guranteed life .

$\therefore \dfrac{\sqrt{n}}{4} = 1.65 \ \text{i.e.,} \ \sqrt{n} = 6.6$

$\therefore n = 44$

11.A normal population has a mean of 0.1 and standard deviation of 2.1 . Find the probability that mean of a sample of size 900 will be negative .

Sol : Given $\mu = 0.1$, $\sigma = 2.1$ and n = 900

The Standard normal variate

$$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{\bar{x} - \mu}{\frac{2.1}{\sqrt{900}}} = \frac{\bar{x} - 0.1}{0.07}$$

$\therefore \bar{x} = 0.1 + 0.007$ z   where z $\sim$ N (0 ,1)

$\therefore$ The required probability , that the sample mean is negative is given by

$$P(\bar{x} < 0) = P (0.1 + 0.07 z < 0)$$
$$= P (0.07 z < - 0.1)$$
$$= P \left(z < \left(\frac{-0.1}{0.07}\right)\right)$$
$$= P (z < -1.43)$$
$$= 0.50 - P (0 < z < 1.43)$$
$$= 0.50 - 0.4236 = 0.0764$$

12.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 $\bar{\bar{x}}$ is used as a point estimator to the true average repair costs , with what confidence we can assert that the maximum error doesn't exceed Rs 10.

Sol : Size of a random sample , n = 80

The mean of random sample , $\bar{x}$ = Rs 472.36

Standard deviation , σ = Rs 62.35

Maximum error of estimate , E_{max} = Rs 10

We have $E_{max} = Z\alpha_{/2} . \frac{\sigma}{\sqrt{n}}$

i.e., $Z\alpha_{/2} = \frac{E_{max} . \sqrt{n}}{\sigma} = \frac{10 \sqrt{80}}{62.35} = \frac{89.4427}{62.35} = 1.4345$

$\therefore Z\alpha_{/2} = 1.43$

The area when z = 1.43 from tables is 0.4236

$\therefore \frac{\alpha}{2} = 0.4236$ i.e ., $\alpha = 0.8472$

$\therefore$ confidence = (1- α) 100% = 84.72 %

Hence we are 84.72% confidence that the maximum error is Rs. 10

13.If we can assert with 95% that the maximum error is 0.05 and P = 0.2 find the size of the sample.

Sol : Given P = 0.2 , E = 0.05

We have Q = 0.8 and $Z\alpha_{/2} = 1.96$ (5% LOS)

We know that maximum error , $E = Z\alpha_{/2} \sqrt{\frac{PQ}{n}}$

$\Rightarrow 0.05 = 1.96 \sqrt{\frac{0.2 \times 0.8}{n}}$

$$\Rightarrow \text{Sample size}, \text{n} = \frac{0.2 \times 0.8 \times (1.96)^2}{(0.05)^2} = 246$$

14.The mean and standard deviation of a population are 11,795 and 14,054 respectively What can one assert with 95 % confidence about the maximum error if $\bar{x}$ = 11,795 and n = 50. And also construct 95% confidence interval for true mean .

Sol: Here mean of population , $\mu = 11795$

S.D of population , $\sigma = 14054$

$\bar{x} = 11795$

$n = $ sample size $ = 50$, maximum error $= Z\alpha_{/2} \cdot \frac{\sigma}{\sqrt{n}}$

$Z\alpha_{/2}$ for 95% confidence $= 1.96$

Max. error $, E = Z\alpha_{/2} \cdot \frac{\sigma}{\sqrt{n}} = 1.96 \cdot \frac{14054}{\sqrt{50}} = 3899$

$\therefore$ Confidence interval $= (\bar{x} - Z\alpha_{/2} \cdot \frac{\sigma}{\sqrt{n}} , \bar{x} + Z\alpha_{/2} \cdot \frac{\sigma}{\sqrt{n}})$

$= (11795\text{-}3899, 11795\text{+}3899)$

$= (7896, 15694)$

15.Find 95% confidence limits for the mean of a normally distributed population from which the following sample was taken 15, 17 , 10 ,18 ,16 ,9, 7, 11, 13 ,14.

Sol: We have $\bar{x} = \frac{15+17+10+18+16+9+7+11+13+14}{10} = 13$

$S^2 = \sum \frac{(x_i - \bar{x})^2}{n-1}$

$= \frac{1}{9} [(15-13)^2 + (15-13)^2 + (15-13)^2 + (15-13)^2 + (15-13)^2 +$
$(15-13)^2 + (15-13)^2 + (15-13)^2 + (15-13)^2 + (15-13)^2]$

$= \frac{40}{3}$

Since $Z\alpha_{/2} = 1.96$, we have

$Z\alpha_{/2} \cdot \frac{s}{\sqrt{n}} = 1.96 \cdot \frac{\sqrt{40}}{\sqrt{10} \cdot \sqrt{3}} = 2.26$

$\therefore$ Confidence limits are $\bar{x} \pm Z\alpha_{/2} \cdot \frac{s}{\sqrt{n}} = 13 \pm 2.26 = (10.74 , 15.26)$

16.A random sample of 100 teachers in a large metropolitan area revealed mean weekly salary of Rs. 487 with a standard deviation Rs.48. With what degree of confidence can we assert that the average weekly of all teachers in the metropolitan area is between 472 to 502 ?

Sol: Given $\mu = 487$, $\sigma = 48$, $n = 100$

$Z = \frac{\bar{\bar{x}} - \mu}{\frac{\sigma}{\sqrt{n}}}$

$= \frac{\bar{x} - 487}{\frac{48}{\sqrt{100}}} = \frac{\bar{x} - 487}{4.8}$

Standard variable corresponding to Rs. 472 is

$Z_1 = \dfrac{472 - 487}{4.8} = -3.125$

Standard vaiable corresponding to Rs. 502

$Z_2 = \dfrac{502 - 487}{4.8} = 3.125$

Let $\bar{x}$ be the mean salary of teacher . Then

$P(472 < \bar{x} < 502) = P(-3.125 < z < 3.125)$

$= 2(0 < z < 3.125)$

$= 2\int_0^{3.125} \emptyset(z)dz$

$= 2(0.4991) = 0.9982$

Thus we can ascertain with 99.82 % confidence

TUTORIAL QUESTIONS

1.If the population is 3,6,9,15,27

a)List all possible samples of size 3 that can be taken without replacement from finite population

b)Calculate the mean of each of the sampling distribution of means

c)Find the standard deviation of sampling distribution of means

2. A population consist 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 standard deviation of the population

       c)The mean of the sampling distribution of means and

    d)   The standard deviation of the sampling distribution of means

3.A random sample of size 100 is taken from a population with $\sigma = 5.1$. Given that the

   sample mean is $\bar{x} = 21.6$ Construct a 95% confidence limits for the population mean .

4.A normal population has a mean of 0.1 and standard deviation of 2.1 . Find the probability

   that mean of a sample of size 900 will be negative .

5.A random sample of size 64 is taken from a normal population with $\mu = 51.4$

   and $\sigma = 6.8$.What is the probability that the mean of the sample will

 a) exceed 52.9

 b) fall between 50.5 and 52.3 c) be less than 50.6.

ASSIGNMENT QUESTIONS

1. A manufacturer claimed that at least 95% of the equipment which he supplied to factory conformed to specifications . An examination of a sample of 200 pieces of equipment revealed that 180 were faulty .Test his claim at 5% an 1% LOS.

2. Write about   i) critical region ii) one tailed and two tailed test

3. Define sample. Explain the different methods that are involved in selecting the sample.

4. Explain about   i) Type I error   ii) Type II error

5.a)Explain the five step procedure for testing of hypothesis

   b) Explain about i) point estimation ii) interval estimation
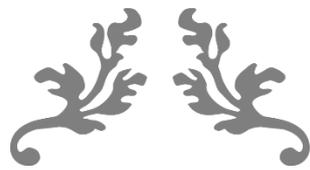
UNIT 5

STATISTICAL INFERENCES

OBJECTIVE

To make inferences about a population from sample data

(large and small samples) using probability theory

OUTCOME

Draw statistical inference using samples of a given size which is

taken from a population and to apply statistical methods for analyzing

experimental data.

STATISTICAL INFERENCES

Large Samples: Let a random sample of size n >30 is defined as large sample.

Applications of Large Samples

Test of Significance of a Single Mean

Let a random sample of size n, $\bar{x}$ be the mean of the sample and μ be the population mean.

1. **Null hypothesis**: H_0: There is no significant difference in the given population mean value say 'μ_0'.

    i.e $H_0 : \mu = \mu_0$

2. **Alternative hypothesis**: H_1 :There is some significant difference in the given population mean value.
    i.e
    a)$H_1 : \mu \neq \mu_0$ (Two –tailed)
    b) $H_1 : \mu > \mu_0$ (Right one tailed)
    c) $H_1 : \mu > \mu_0$ (Left one tailed)
3. **Level of significance**: Set the LOS α
4. **Test Statistic**: $z_{cal} = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$ (OR) $z_{cal} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$
5. **Decision /conclusion** : If z_{cal} value $< z_\alpha$ value , accept H_0 otherwise reject H_0

CRITICAL VALUES OF Z

LOS $\propto$	1%	5%	10%
$\mu \neq \mu_0$	/Z/>2.58	/Z/>1.96	/Z/>1.645
$\mu > \mu_0$	Z>2.33	z>1.645	Z>1.28
$\mu < \mu_0$	Z<-2.33	Z<-1.645	Z<-1.28

NOTE: Confidence limits for the mean of the population corresponding to the given sample.

$\mu = \bar{X} \pm Z\propto_{/2}$ (S.E of $\bar{X}$) i.e,

$\mu = \bar{X} \pm Z\propto_{/2} \left(\frac{\sigma}{\sqrt{n}}\right)$ (or) $\mu = \bar{X} \pm Z\propto_{/2} \left(\frac{\varepsilon}{\sqrt{n}}\right)$

2. Test of Significance for Difference of Means of two Large Samples

Let $\overline{x_1}$ & $\overline{x_2}$ be the means of the samples of two ramdom sizes n_1 & n_2 drawn from two populations having means μ_1 & μ_2 and SD's σ_1 & σ_2

i) **Null hyopothesis**: $H_0 : \mu_1 = \mu_2$

ii) **Alternative hypothesis** :. a) $H_1 : \mu_1 \neq \mu_2$ (Two Tailed)

b) $H_1 : \mu_1 < \mu_2$ (Left one tailed)
c) $H_1 : \mu_1 > \mu_2$ (Right one tailed)

iii) **Level of Significance**: Set the LOS α

iv) **Test Statistic** : $Z_{cal} = \frac{(\bar{x}_1 - \bar{x}_2) - \delta}{SE \, of(\bar{x}_1 - \bar{x}_2)} = \frac{(\bar{x}_1 - \bar{x}_2) - \delta}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$

Where $\delta = \mu_1 - \mu_2$ (where given constant)
Other wise $\delta = \mu_1 - \mu_2 = 0$

$Z_{cal} = \frac{\bar{x}_1 - \bar{x}_2 - \delta}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$ if $\sigma_1^2 = \sigma_2^2 = \sigma^2$ then $Z_{cal} = \frac{\bar{X}_1 - \bar{X}_2}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$

Critical value of Z from normal table at the LOS α

v)   **Decision:** If $|Z_{cal}| < Z_{tab}$, accept H$_0$ otherwise reject H$_0$

CRITICAL VALUES OF Z

LOS ∝	1%	5%	10%
μ≠ μ$_0$	/Z/>2.58	/Z/>1.96	/Z/>1.645
μ> μ$_0$	Z>2.33	z>1.645	Z>1.28
μ< μ$_0$	Z<-2.33	Z<-1.645	Z<-1.28

NOTE: Confidence limits for difference of means

$$\mu_1 - \mu_2 = (\bar{X}_1 - \bar{X}_2) \pm z_{\propto/2}[S.E \, of \, (\bar{X}_1 - \bar{X}_2)]$$

$$= (\bar{X}_1 - \bar{X}_2) \pm z_{\propto/2}\left[\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_1}}\right]$$

3. Test of Significance for Single Proportions

Suppose a random sample of size n has a sample proportion p of members possessing a certain attribute (proportion of successes). To test the hypothesis that the proportion P in the population has a specified value P$_0$.

i)   **Null hyopthesis** : $H_0: P = P_0$
ii)   **Alternative hypothesis** : a) H$_1$: P≠ P_0(Two Tailed test)
    b) H_1 : $P < P_0$ (Left one- tailed)
    c) H_1 : P $> P_0$ (Right one tailed)
iii) **Test statistic** :$Z_{cal} = \frac{p - P}{\sqrt{\frac{PQ}{n}}}$ when P is the Population proportion $Q = 1 - P$

iv) At specified LOS ∝, critical value of Z
v) **Decision**: If $|z_{cal}| < Z_{tab}$, accept H$_0$ otherwise reject H$_0$

CRITICAL VALUES OF Z

LOS ∝	1%	5%	10%
μ≠ μ$_0$	/Z/>2.58	/Z/>1.96	/Z/>1.645
μ> μ$_0$	Z>2.33	z>1.645	Z>1.28
μ< μ$_0$	Z<-2.33	Z<-1.645	Z<-1.28

NOTE : Confidence limits for population proportion

$$P = P \pm Z_{\frac{\propto}{2}}(S \, E \, of \, P)$$

$$= P \pm Z_{\frac{\alpha}{2}} \left(\sqrt{\frac{pq}{n}} \right)$$

4. Test for Equality of Two Proportions (Populations)

Let p_1 and p_2 be the sample proportions in two large random samples of sizes n_1 & n_2 drawn from two populations having proportions P_1 & P_2

i) **Null hyopothesis** : $H_0: P_1 = P_2$

ii) **Alternative hypothesis** : a) $H_1 : P_1 \neq P_2$ (Two Tailed)

   b) $H_1 : P_1 < P_2$ (Left one tailed)

   c) $H_1 : P_1 > P_2$ (Right one tailed)

iii) **Test statistic** : $Z_{cal} = \dfrac{(P_1 - P_2) - (P_1 - P_2)}{\sqrt{\frac{P_1 Q_1}{n_1} + \frac{P_1 Q_1}{n_2}}}$ if $(P_1 - P_2)$ is given.

If given only sample proportions then

$$Z_{cal} = \dfrac{p_1 - p_2}{\sqrt{\frac{P_1 q_1}{n_1} + \frac{P_2 q_1}{n_2}}} \quad \text{where} \quad p_1 = \frac{x_1}{n_1} \text{ \& } p_1 = \frac{x_2}{n_2}$$

OR

$$Z_{cal} = \dfrac{p_1 - p_2}{\sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \quad \text{Where} \quad p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{x_1 + x_2}{n_1 + n_2} \text{ and } q = 1 - p$$

iv) At specified LOS $\propto$ critical value of 'Z'

v) **Decision**: If $|Z_{cal}| < Z_{Tab}$, accept H_0 otherwise reject H_0

CRITICAL VALUES OF Z

LOS $\propto$	1%	5%	10%
$\mu \neq \mu_0$	/Z/>2.58	/Z/>1.96	/Z/>1.645
$\mu > \mu_0$	Z>2.33	z>1.645	Z>1.28
$\mu < \mu_0$	Z<-2.33	Z<-1.645	Z<-1.28

NOTE: Confidence limits for difference of population proportions

$$P_1 - P_2 = (p_1 - p_2) \pm Z_{\frac{\alpha}{2}} (S.E \text{ of } P_1 - P_2)$$

Problems:

1. **A sample of 64 students have a mean weight of 70 kgs . Can this be regarded as asample mean from a population with mean weight 56 kgs and standard deviation 25 kgs.**

   Sol : Given $\bar{x}$ = mean of he sample = 70 kgs

   μ = Mean of the population = 56 kgs

   σ = S.D of population = 25 kgs

and n = Sample size = 64

i)   Null Hypothesis H_0 : A Sample of 64 students with mean weight 70 kgs be regarded as a sample from a population with mean weight 56 kgs and standard deviation 25 kgs. i.e., $H_0 : \mu = 70$ kgs

ii)   Alternative Hypothesis H_1 : Sample cannot be regarded as one coming from the population . i.e., $H_1 : \mu \neq 70$ kgs (Two –tailed test)

iii)   Level of significance : $\alpha = 0.05$ ($Z_\alpha = 1.96$)

iv)   Test Statistic : $Z_{cal} = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{70 - 56}{\frac{25}{\sqrt{64}}} = 4.48$

v)   Conclusion: Since $|Z_{cal}|$ $value > Z_\alpha$ value , we reject H_0

     $\therefore$ Sample cannot be regarded as one coming from the population

2. **In a random sample of 60 workers , the average time taken by them to get to work is 33.8 minutes with a standard deviation of 6.1 minutes . Can we reject the null hypothesis μ = 32.6 in favor of alternative null hypothesis μ > 32.6 at α = 0.05 LOS**

     Sol : Given n = 60 , $\bar{x}$ = 33.8 , μ = 32.6 and σ = 6.1

i)   Null Hypothesis $H_0 : \mu = 32.6$

ii)   Alternative Hypothesis $H_1 : \mu > 32.6$ (Right one tailed test)

iii)   Level of significance : $\alpha = 0.01$ ($Z_\alpha = 2.33$)

iv)   Test Statistic : $Z_{cal} = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{33.8 - 32.6}{\frac{6.1}{\sqrt{60}}} = \frac{1.2}{0.7875} = 1.5238$

v)   Conclusion: Since Z_{cal} $value < Z_\alpha$ value , we accept H_0

3. **A sample of 400 items is taken from a population whose standard deviation 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 limits for the population .**

     Sol : Given n = 400 , $\bar{x}$ = 40 , μ = 38 and σ = 10

i)   Null Hypothesis $H_0 : \mu = 38$

ii)   Alternative Hypothesis $H_1 : \mu \neq 38$ (Two –tailed test)

iii)   Level of significance : $\alpha = 0.05$ ($Z_\alpha = 1.96$)

iv)   Test Statistic : $Z_{cal} = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{38 - 40}{\frac{10}{\sqrt{400}}} = \frac{-2}{0.5} = -4$

v)   Conclusion: Since $|Z_{cal}|$ $v\_lue > Z_\alpha$ value , we reject H_0

     i.e., the sample is not from the population whose is 38.

     $\therefore$ 95% confidence interval is $\left(\bar{x} - 1.96 . \frac{\sigma}{\sqrt{n}} , \bar{x} + 1.96 . \frac{\sigma}{\sqrt{n}} \right)$

     i.e., $\left(40 - \frac{1.96(10)}{\sqrt{400}} , 40 + \frac{1.96(10)}{\sqrt{400}} \right)$

     $= \left(40 - \frac{1.96(10)}{20} , 40 + \frac{1.96(10)}{20} \right)$

     $= (40 - 0.98 , 40 + 0.98)$

     $= (39.02 , 40.98)$

4. **An insurance agent has claimed that the average age of policy holders who issue through him is less than the average for all agents which is 30.5. A random sample of 100 policy holders who had issued through him gave the following age distribution .**

Age	16-20	21-25	26-30	31-35	36-40
No# of persons	12	22	20	30	16

Calculate the arithmetic mean and standard deviation of this distribution and use these values to test his claim at 5% los.

Sol : Take A = 28 where A – Assumed mean

$$d_i = x_i - A$$

$$\bar{x} = A + \frac{h \, \Sigma f_i d_i}{N}$$

$$= 28 + \frac{5 \times 16}{100} = 28.8$$

S.D : $S = h \sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2} = 5. \sqrt{\frac{164}{100} - \left(\frac{16}{100}\right)^2} = 6.35$

i) Null Hypothesis H_0 : The sample is drawn from population with mean μ
ii) i.e., $H_0 : \mu = 30.5$ years
iii) Alternative Hypothesis $H_1 : \mu < 30.5$ (Left one –tailed test)
iv) Level of significance : $\alpha = 0.05$ ($Z_\alpha = 1.645$)
v) Test Statistic : $Z_{cal} = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{28.8 - 30.5}{\frac{6.35}{\sqrt{100}}} = -2.677$
vi) Conclusion: Since $|Z_{cal}| \, value > Z_\alpha$ value , we reject H_0
i.e., the sample is not drawn from the population with $\mu = 30.5$ years .

5. **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 los?**
Sol : Given n = 36 , $\bar{x}$ =11 , $\mu = 10$ and $\sigma = \sqrt{16} = 4$
i) Null Hypothesis $H_0 : \mu = 10$
ii) Alternative Hypothesis $H_1 : \mu < 10$ (Left one –tailed test)
iii) Level of significance : $\alpha = 0.05$ ($Z_\alpha = 1.645$)
iv) Test Statistic : $Z_{cal} = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{11 - 10}{\frac{4}{\sqrt{36}}} = \frac{6}{4} = 1.5$
v) Conclusion: Since $|Z_{cal}| \, value < Z_\alpha$ value , we accept H_0

6. **The means of two large samples of sizes 1000 and 2000 members are 67.5 inches and 68 inches respectively . Can the samples be regarded as drawn from the same population of S.D 2.5 inches.**
Sol: Let μ_1 and μ_2 be the means of the two populations

Given $n_1 = 1000$, $n_2 = 2000$ and $\overline{x}_1 = 67.5$ inches , $\overline{x}_2 = 68$ inches

Population S.D, $\sigma = 2.5$ inches

i) Null Hypothesis H_0 :The samples have been drawn from the same population of S.D 2.5 inches

i.e., $H_0 : \mu_1 = \mu_2$

ii) Alternative Hypothesis $H_1 : \mu_1 \neq \mu_2$ (Two – Tailed test)

iii) Level of significance : $\alpha = 0.05$ ($Z_\alpha = 1.96$)

iv) Test Statistic : $Z_{cal} = \dfrac{\overline{X}_1 - \overline{X}_2}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \dfrac{67.5 - 68}{\sqrt{(2.5)^2\left(\frac{1}{1000} + \frac{1}{2000}\right)}} = \dfrac{-0.5}{0.0968} = -5.16$

v) Conclusion: Since $|Z_{cal}|\ value > Z_\alpha$ value , we reject H_0

Hence , we conclude that the samples are not drawn from the same population of S.D 2.5 inches.

7. **Samples of students were drawn from two universities and from their weights in kilograms , mean and standard deviations are calculated and shown below. Make a large sample test to test the significance of the difference between the means.**

	Mean	S .D	Size of the sample
University A	55	10	400
University B	57	15	100

Sol: Let μ_1 and μ_2 be the means of the two populations

Given $n_1 = 400$, $n_2 = 100$ and $\overline{x}_1 = 55$ kgs , $\overline{x}_2 = 57$ kgs

$\sigma_1 = 10$ and $\sigma_2 = 15$

i) Null Hypothesis $H_0 : \mu_1 = \mu_2$

ii) Alternative Hypothesis $H_1 : \mu_1 \neq \mu_2$ (Two – Tailed test)

iii) Level of significance : $\alpha = 0.05$ ($Z_\alpha = 1.96$)

iv) Test Statistic : $Z_{cal} = \dfrac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \dfrac{55 - 57}{\sqrt{\frac{10^2}{400} + \frac{15^2}{100}}} = \dfrac{-2}{\sqrt{\frac{1}{4} + \frac{9}{4}}} = -1.26$

v) Conclusion: Since $|Z_{cal}|\ value < Z_\alpha$ value , we accept H_0

Hence , we conclude that there is no significant difference between the means

8. **The average marks scored by 32 boys is 72 with a S.D of 8 . While that for 36 girls is 70 with a S.D of 6. Does this data indicate that the boys perform better than girls at 5% los ?**

Sol: Let μ_1 and μ_2 be the means of the two populations

Given $n_1 = 32$, $n_2 = 36$ and $\overline{x}_1 = 72$, $\overline{x}_2 = 70$

$\sigma_1 = 8$ and $\sigma_2 = 6$

i) Null Hypothesis $H_0 : \mu_1 = \mu_2$

ii) Alternative Hypothesis $H_1 : \mu_1 > \mu_2$ (Right One Tailed test)

iii) Level of significance : $\alpha = 0.05$ ($Z_\alpha = 1.645$)

iv) Test Statistic : $Z_{cal} = \dfrac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \dfrac{72 - 70}{\sqrt{\frac{8^2}{32} + \frac{6^2}{36}}} = \dfrac{2}{\sqrt{2+1}} = 1.1547$

v) Conclusion: Since $|Z_{cal}|\ value < Z_\alpha$ value , we accept H_0

Hence , we conclude that the performance of boys and girls is the same

9. **A sample of the height of 6400 Englishmen has a mean of 67.85 inches and a S.D of 2.56 inches while another sample of heights of 1600 Austrians has a mean of 68.55 inches and S.D of 2.52 inches. Do the data indicate that Austrians are on the average taller than the Englishmen ? (Use α as 0.01)**

Sol : Let μ_1 and μ_2 be the means of the two populations

Given $n_1 = 6400$, $n_2 = 1600$ and $\bar{x}_1 = 67.85$, $\bar{x}_2 = 68.55$

$\sigma_1 = 2.56$ and $\sigma_2 = 2.52$

i)   Null Hypothesis $H_0 : \mu_1 = \mu_2$

ii)   Alternative Hypothesis $H_1 : \mu_1 < \mu_2$ (Left One Tailed test)

iii)   Level of significance : $\alpha = 0.01$ $(Z_\alpha = -2.33)$

iv)   Test Statistic : $Z_{cal} = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} = \dfrac{67.85 - 68.55}{\sqrt{\dfrac{2.56^2}{6400} + \dfrac{2.52^2}{1600}}}$

$$= \dfrac{67.85 - 68.55}{\sqrt{\dfrac{6.5536}{6400} + \dfrac{6.35}{1600}}}$$

$$= \dfrac{-0.7}{\sqrt{0.001 + 0.004}} = \dfrac{-0.7}{0.0707} \text{ - } 9.9$$

v)   Conclusion: Since $|Z_{cal}|$ value $> Z_\alpha$ value , we reject H_0

Hence , we conclude that Australians are taller than Englishmen.

10. **At a certain large university a sociologist speculates that male students spend considerably more money on junk food than female students. To test her hypothesis the sociologist randomly selects from records the names of 200 students . Of thee , 125 are men and 75 are women . The mean of the average amount spent on junk food per week by the men is Rs. 400 and S.D is 100. For the women the sample mean is Rs. 450 and S.D is 150. Test the hypothesis at 5 % los ?**

Sol: Let μ_1 and μ_2 be the means of the two populations

Given $n_1 = 125$, $n_2 = 75$ and $\bar{x}_1 = $ Mean of men $= 400$, $\bar{x}_2 = $ Mean of women $= 450$

$\sigma_1 = 100$ and $\sigma_2 = 150$

i)   Null Hypothesis $H_0 : \mu_1 = \mu_2$

ii)   Alternative Hypothesis $H_1 : \mu_1 > \mu_2$ (Right One Tailed test)

iii)   Level of significance : $\alpha = 0.05$ $(Z_\alpha = 1.645)$

iv)   Test Statistic : $Z_{cal} = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} = \dfrac{400 - 450}{\sqrt{\dfrac{100^2}{125} + \dfrac{150^2}{75}}}$

$$= \dfrac{-50}{\sqrt{80 + 300}}$$

$$= \dfrac{-50}{\sqrt{380}} = \dfrac{-50}{19.49} = \text{ - } 2.5654$$

v)   Conclusion: Since Z_{cal} value $< Z_\alpha$ value , we accept H_0

Hence , we conclude that difference between the means are equal

11. **The research investigator is interested in studying whether there is a significant difference in the salaries of MBA grads in two cities. A random sample of size 100 from city A yields an average income of Rs. 20,150 . Another random sample**

of size 60 from city B yields an average income of Rs. 20,250. If the variance are given as $\sigma_1{}^2 = 40,000$ and

$\sigma_2{}^2 = 32,400$ respectively . Test the equality of means and also construct 95% confidence limits.

Sol: Let μ_1 and μ_2 be the means of the two populations

Given $n_1 = 100$, $n_2 = 60$ and $\bar{x}_1 =$ Mean of city A = 20,150 , $\bar{x}_2 =$ Mean of city B = 20,250

$\sigma_1{}^2 = 40,000$ and $\sigma_2{}^2 = 32,400$

    i)   Null Hypothesis $H_0 : \mu_1 = \mu_2$

    ii)   Alternative Hypothesis $H_1 : \mu_1 \neq \mu_2$ (Two -Tailed test)

    iii)   Level of significance : $\alpha = 0.05$ ($Z_\alpha = 1.96$)

    iv)   Test Statistic : $Z_{cal} = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\dfrac{\sigma_1{}^2}{n_1} + \dfrac{\sigma_2{}^2}{n_2}}} = \dfrac{20{,}150 - 20{,}250}{\sqrt{\dfrac{40000}{100} + \dfrac{32400}{60}}}$

$$= \frac{100}{\sqrt{400 + 540}}$$

$= \dfrac{100}{30.66} = 3.26$

    v)   Conclusion: Since Z_{cal} value $> Z_\alpha$ value , we reject H_0

Hence , we conclude that there is a significant difference in the salaries of MBA grades two cities.

$\therefore$ 95% confidence interval is $\mu_1 - \mu_2 = (\bar{x_1} - \bar{x_2}) \pm 1.96 \sqrt{\dfrac{\sigma_1{}^2}{n_1} + \dfrac{\sigma_2{}^2}{n_2}}$

$= (20{,}150 - 20{,}250)) \pm 1.96 \sqrt{\dfrac{40000}{100} + \dfrac{32400}{60}} = (39.90,\ 160.09)$

12. **A die was thrown 9000 times and of these 3220 yielded a 3 or 4. Is this consistent with the hypothesis that the die was unbiased?**

Sol : Given n = 9000

P = Population of proportion of successes

   = P(getting a 3 or 4) $= \dfrac{1}{6} + \dfrac{1}{6} = \dfrac{2}{6} = \dfrac{1}{3}$ 0.3333

Q = 1- P = 0.6667

P = Proportion of successes of getting 3 or 4 in 9000 times $= \dfrac{3220}{9000} = 0.3578$

i)   Null Hypothesis H_0 : The die is unbiased

        i.e., H_0 : P = 0.33

ii) Alternative Hypothesis H_1 : The die is biased

       i.e., H_1 : P $\neq$ 0.33 (Two –Tailed test)

iii) Level of Significance : $\alpha = 0.05$ ($Z_\alpha = 1.96$)

iv) Test Statistic : $Z_{cal} = \dfrac{p - P}{\sqrt{\dfrac{PQ}{n}}} = \dfrac{0.3578 - 0.3333}{\sqrt{\dfrac{(0.3333)(0.6667)}{9000}}} = 4.94$

v)Conclusion: Since Z_{cal} value $> Z_\alpha$ value , we reject H_0

Hence , we conclude that the die is biased.

13. In a random sample of 125 cool drinkers , 68 said they prefer thumsup to Pepsi . Test the null hypothesis P = 0.5 against the alternative hypothesis hypothesis P > 0.5?

Sol : Given n = 125 , x = 68 and $p = \frac{x}{n} = \frac{68}{125} = 0.544$

i)   Null Hypothesis H_0 : P = 0.5

ii)   Alternative Hypothesis H_1 : P > 0.5(Right One Tailed test)

iii)   Level of Significance : $\alpha = 0.05$ ($Z_\alpha = 1.645$)

iv)   Test Statistic : $Z_{cal} = \frac{p-P}{\sqrt{\frac{PQ}{n}}} = \frac{0.544-0.5}{\sqrt{\frac{(0.5)(0.5)}{125}}} = 0.9839$

v)   Conclusion: Since Z_{cal} value $<$ Z_α value , we accept H_0

14. A manufacturer claimed that at least 95% of the equipment which he supplied to a factory conformed to specifications . An experiment of a sample of 200 piece of equipment revealed that 18 were faulty .Test the claim at 5% los ?

Sol : Given n = 200

Number of pieces confirming to specifications = 200-18 = 182

$\therefore$ p = Proportion of pieces confirming to specification $= \frac{182}{200} = 0.91$

P = Population proportion $= \frac{95}{100} = 0.95$

i)   Null Hypothesis H_0 : P = 0.95

ii)   Alternative Hypothesis H_1 : P < 0.95(Left One Tailed test)

iii)   Level of Significance : $\alpha = 0.05$ ($Z_\alpha = -1.645$)

iv)   Test Statistic : $Z_{cal} = \frac{p-P}{\sqrt{\frac{PQ}{n}}} = \frac{0.91-0.95}{\sqrt{\frac{0.95 \times 0.05}{200}}} = - 2.59$

v)   Conclusion: We reject H_0

Hence , we conclude that the manufacturer's claim is rejected.

15. Among 900 people in a state 90 are found to be chapatti eaters . Construct 99% confidence interval for the true proportion and also test the hypothesis for single proportion ?

Sol: Given x = 90 , n = 900

$\therefore p = \frac{x}{n} = \frac{90}{100} = \frac{1}{10} = 0.1$

And q = 1- p= 0.9

Now $\sqrt{\frac{pq}{n}} = \sqrt{\frac{(0.1)(0.9)}{900}} = 0.01$

Confidence interval is $P = p \pm Z_{\frac{\alpha}{2}} \left(\sqrt{\frac{pq}{n}} \right)$

i.e., (0.1- 0.03 , 0.1 + 0.03)

= (0.07 , 0.13)

i)   Null Hypothesis H_0 : P = 0.5

ii)   Alternative Hypothesis H_1 : P $\neq$ 0.5(Two Tailed test)

iii)   Level of Significance : $\alpha = 0.01$ ($Z_\alpha = 2.58$)

iv)   Test Statistic : $Z_{cal} = \frac{p-P}{\sqrt{\frac{PQ}{n}}} = \frac{0.1-0.5}{\sqrt{\frac{0.5 \times 0.5}{900}}} = -24.39$

v)   Conclusion: Since $|Z_{cal}| value > Z_\alpha$ value , we reject H_0

16. **Random samples of 400 men and 200 women in a locality were asked whether they would like to have a bus stop a bus stop near their residence . 200 men and 40 women in favor of the proposal . Test the significance between the difference of two proportions at 5% los ?**

Sol: Let P_1 and P_2 be the population proportions in a locality who favor the bus stop

Given n_1 = Number of men = 400

n_2 = number of women = 200

x_1 = Number of men in favor of the bus stop = 200

x_2 = Number of women in favor of the bus stop 40

$\therefore p_1 = \frac{x_1}{n_1} = \frac{200}{400} = \frac{1}{2}$ and $p_2 = \frac{x_2}{n_2} = \frac{40}{200} = \frac{1}{5}$

i)   Null Hypothesis $H_0 : P_1 = P_2$

ii)   Alternative Hypothesis $H_1 : P_1 \neq P_2$ (Two Tailed test)

iii)   Level of Significance : $\alpha = 0.05$ $(Z_\alpha = 1.96)$

iv)   Test Statistic : $Z_{cal} = \dfrac{p_1 - p_2}{\sqrt{pq\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$

We have p $= \dfrac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \dfrac{x_1 + x_2}{n_1 + n_2} = \dfrac{200+40}{400+200} = \dfrac{240}{600} = \dfrac{2}{5}$

q = 1- p $= \dfrac{3}{5}$

$= \dfrac{0.5-0.2}{\sqrt{(0.4)(0.6)\left(\frac{1}{400}+\frac{1}{200}\right)}} = 7.07$

v)   Conclusion: Since $|Z_{cal}|$value $> Z_\alpha$ value , we reject H_0

Hence we conclude that there is difference between the men and women in their attitude towards the bus stop near their residence.

17. **A machine puts out 16 imperfect articles in a sample of 500 articles . After the machine is overhauled it puts out 3 imperfect articles in a sample of 100 articles . Has the machine is improved ?**

Sol : Let P_1 and P_2 be the proportions of imperfect articles in the proportion of articles manufactured by the machine before and after overhauling , respectively.

Given n_1 = Sample size before the machine overhauling = 500

n_2 = Sample size after the machine overhauling = 100

x_1 = Number of imperfect articles before overhauling = 16

x_2 = Number of imperfect articles after overhauling = 3

$\therefore p_1 = \frac{x_1}{n_1} = \frac{16}{500} = 0.032$ and $p_2 = \frac{x_2}{n_2} = \frac{3}{100} = 0.03$

i)   Null Hypothesis $H_0 : P_1 = P_2$

ii)   Alternative Hypothesis $H_1 : P_1 > P_2$ (Left one Tailed test)

iii)   Level of Significance : $\alpha = 0.05$ $(Z_\alpha = 1.645)$

iv)   Test Statistic : $Z_{cal} = \dfrac{p_1 - p_2}{\sqrt{pq\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$

We have p $= \dfrac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \dfrac{x_1 + x_2}{n_1 + n_2} = \dfrac{16+3}{500+100} = \dfrac{19}{600} = 0.032$

q = 1- p = 0.968

$= \dfrac{0.032-0.03}{\sqrt{(0.032)(0.968)\left(\frac{1}{500}+\frac{1}{100}\right)}}$

$$\frac{0.002}{0.019} = 0.104$$

v)   Conclusion: Since $|Z_{cal}|$value $< Z_\alpha$ value , we accept H_0

Hence we conclude that the machine has improved.

18. **In an investigation on the machine performance the following results are obtained .**

	No# of units inspected	No# of defectives
Machine 1	375	17
Machine 2	450	22

Test whether there is any significant performance of two machines at $\alpha = 0.05$

Sol: Let P_1 and P_2 be the proportions of defective units in the population of units inspected in machine 1 and Machine 2 respectively.

Given n_1 = Sample size of the Machine 1 = 375

n_2 = Sample size of the Machine 2 = 450

x_1 = Number of defectives of the Machine 1 = 17

x_2 = Number of defectives of the Machine 2 = 22

$\therefore p_1 = \frac{x_1}{n_1} = \frac{17}{375} = 0.045$ and $p_2 = \frac{x_2}{n_2} = \frac{22}{450} = 0.049$

i)   Null Hypothesis $H_0 : P_1 = P_2$

ii)   Alternative Hypothesis $H_1 : P_1 \neq P_2$ (Two Tailed test)

iii)   Level of Significance : $\alpha = 0.05$ $(Z_\alpha = 1.96)$

iv)   Test Statistic : $Z_{cal} = \frac{p_1 - p_2}{\sqrt{pq\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$

We have p $= \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{17+22}{375+450} = \frac{39}{825} = 0.047$

q = 1- p = 1- 0.047 = 0.953

$= \frac{0.045 - 0.049}{\sqrt{(0.047)(0.953)\left(\frac{1}{375}+\frac{1}{450}\right)}}$

$= - 0.267$

v)   Conclusion: Since $|Z_{cal}|$value $< Z_\alpha$ value , we accept H_0

Hence we conclude that there is no significant difference in performance of machines.

19. **A cigarette manufacturing firm claims that its brand A line of cigarettes outsells its**

brand B by 8% . If it is found that 42 out of 200 smokers prefer brand A and 18 out of another sample of 100 smokers prefer brand B . Test whether 8% difference is a valid claim?

Sol: Given $n_1 = 200$

$n_2 = 100$

x_1 = Number of smokers preferring brand A= 42

x_2 = Number of smokers preferring brand B = 18

$\therefore p_1 = \frac{x_1}{n_1} = \frac{42}{200} = 0.21$ and $p_2 = \frac{x_2}{n_2} = \frac{18}{100} = 0.18$

and $P_1 - P_2 = 8\% = 0.08$

i)   Null Hypothesis $H_0 : P_1 - P_2 = 0.08$

ii)   Alternative Hypothesis $H_1 : P_1 - P_2 \neq 0.08$ (Two Tailed test)

iii)   Level of Significance : $\alpha = 0.05$ ($Z_\alpha = 1.96$)

iv)   Test Statistic : $Z_{cal} = \dfrac{(p_1 - p_2)-(P_1 - P_2)}{\sqrt{pq\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$

We have p $= \dfrac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \dfrac{x_1 + x_2}{n_1 + n_2} = \dfrac{42+18}{200+100} = \dfrac{60}{300} = 0.2$

$q = 1- p = 1- 0.2 = 0.8$

$Z_{cal} = \dfrac{(0.21-0.18)-0.08}{\sqrt{(0.2)(0.8)\left(\frac{1}{200}+\frac{1}{100}\right)}}$

$\qquad = \dfrac{-0.05}{0.0489} = -1.02$

v)   Conclusion: Since $|Z_{cal}| value < Z_\alpha$ value , we accept H_0

Hence we conclude that 8% difference in the sale of two brands of cigarettes is a valid claim.

20. **In a city A , 20% of a random sample of 900 schoolboys has a certain slight physical defect . In another city B ,18.5% of a random sample of 1600 school boys has the same defect . Is the difference between the proportions significant at 5% los?**

Sol: Given $n_1 = 900$

$n_2 = 1600$

$x_1 = 20\%$ of $900 = 180$

$x_2 = 18.5\%$ of $1600 = 296$

$\therefore p_1 = \dfrac{x_1}{n_1} = \dfrac{180}{900} = 0.2$ and $p_2 = \dfrac{x_2}{n_2} = \dfrac{296}{1600} = 0.185$

i)   Null Hypothesis $H_0 : P_1 = P_2$

ii)   Alternative Hypothesis $H_1 : P_1 \neq P_2$ (Two Tailed test)

iii)   Level of Significance : $\alpha = 0.05$ ($Z_\alpha = 1.96$)

iv)   Test Statistic : $Z_{cal} = \dfrac{(p_1 - p_2)}{\sqrt{pq\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$

We have p $= \dfrac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \dfrac{x_1 + x_2}{n_1 + n_2} = \dfrac{180+296}{900+1600} = \dfrac{476}{2500} = 0.19$

$q = 1- p = 1- 0.19 = 0.81$

$Z_{cal} = \dfrac{0.2-0.185}{\sqrt{(0.19)(0.81)\left(\frac{1}{900}+\frac{1}{1600}\right)}}$

$\qquad = \dfrac{-0.015}{0.01634} = -0.918$

v)   Conclusion: Since $|Z_{cal}| value < Z_\alpha$ value , we accept H_0

Hence we conclude that there is no significant difference between the proportions.

SMALL SAMPLES

Introduction When the sample size n < 30, then if is referred to as small samples. In this sampling distribution in many cases may not be normal ie., we will not be justified in estimating the population parameters as equal to the corresponding sample values.

Degree Of Freedom The number of independent variates which make up the statistic is known as the degrees of freedom (d.f) and it is denoted by ϑ.

For Example**:** If $x_1 + x_2 + x_3 = 50$ and we assign any values to two os the variables (say x_1, x_2), then the values of x$_3$ will be known. Thus, the two variables are free and independent choices for finding the third.

In general, the number of degrees of freedom is equal to the total number of observations less the number of independent constraints imposed on the observations.

For example: in a set of data of n observations, if K is the number of independent constraints then $\vartheta = n - k$

Student's t-Distribution Or t-Distribution

Let $\bar{X}$ be the mean of a random sample of size n, taken from a normal population having the mean μ and the variance σ^2, and sample variance $S^2 = \sum \dfrac{(X_i - \bar{X})^2}{n-1}$, then

$t = \dfrac{\bar{x} - \mu}{S/\sqrt{n}}$ is a random variable having the $t-$ distribution with $\vartheta = n - 1$ degrees of freedom.

Properties of $t-$ Distribution

1. The shape of $t-$distribution is bell shaped, which is similar to that of normal distribution and is symmetrical about the mean.
2. The mean of the standard normal distribution as well as $t-$distribution is zero, but the variance of $t-$distrubution depends upon the parometer ϑ which is called the degrees of freedom.
3. The variance of $t-$distribution exceeds 1, but approaches 1 as $n \to \infty$.

t Distribution

The t-distribution is used when n is **small** and σ is **unknown**.
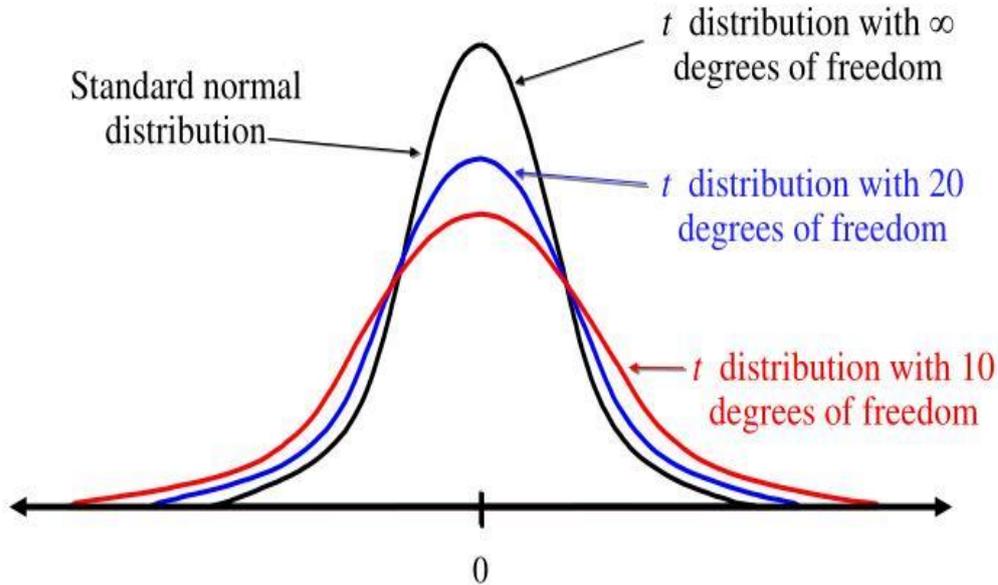
Applications of t – Distributions

1. To test the significance of the sample mean, When population variance is not given:

Let $\bar{x}$ be the mean of the sample and n be the size of the sample 'σ' be the standard deviation of the population and μ be the mean of the population.

Then the student $t-$ distribution is defined by the statistic

$t = \frac{\bar{x}-\mu}{\frac{s}{\sqrt{n-1}}}$ if s is given directly

If $'\sigma'$ is unknown, then $t = \frac{\bar{X}-\mu}{S/\sqrt{n}}$ where

$S^2 = \sum \frac{(X_i-\bar{X})^2}{n-1}$

Note : Confidence limits for mean $\mu = \bar{x} \pm t_\alpha(S/\sqrt{n})$ or $\mu = \bar{x} \pm t_\alpha(S/\sqrt{n-1})$

2. To test the significance of the difference between means of the two independent samples :

To test the significant difference between the sample means $\bar{x}_1$ and $\bar{x}_2$ of two independent samples of sizes n_1 and n_2, with the same variance .

DEPARTMENT OF HUMANITIES & SCIENCES

We use statistic

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{S^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \text{-------(1) where}$$

$\bar{x}_1 = \frac{\sum x_1}{n_1}, \bar{x}_2 = \frac{\sum x_2}{n_2}$ and

$$S^2 = \frac{1}{n_1 + n_2 - 2}\left[\sum(x_1 - \bar{x}_1)^2 + \sum(x_2 - \bar{x}_2)^2\right.$$

OR $\quad S^2 = \frac{1}{n_1 + n_2 - 2}\left[(n_1 s_1^2) + (n_2 s_2^2)\right]$

Where s_1 and s_2 are sample standard deviations.

Note: Confidence limits for difference of means $\quad : \quad \mu_1 - \mu_2 = (\bar{x}_1 - \bar{x}_2)$
$\pm t_\alpha \left(\sqrt{S^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}\right)$

3.Paired t- test (Test the significance of the difference between means of two dependent samples) :

Paired observations arise in many practical situations where each homogenous experimental unit receives both population condition.

For Example: To test the effectiveness of 'drug' some // person's blood pressure is measured before and after the intake of certain drug. Here the individual person is the experimental unit and the two populations are blood pressure "before" and "after" the drug is given

Paired t-test is applied for n paired observations by taking the differences d_1, d_2 ------d_n of the paired data. To test whether the differences d_i from a random sample of a population with mean μ.

$$t = \frac{\bar{d}}{s/\sqrt{n}} \text{ where } \bar{d} = \frac{1}{n}\epsilon \, d_i \text{ and } s^2 = \frac{1}{n-1}\sum(d - \bar{d})^2$$

Problems:

1. **A sample of 26 bulbs gives a mean life of 990 hours with a S.D of 20 hours. The manufacturer claims that the mean life of bulbs is 1000 hours . Is the sample not upto the standard?**

 Sol: Given n = 26

 $\bar{x} = 990$

 $\mu = 1000$ and S.D i.e., s = 20

 i) Null Hypothesis : $H_0 : \mu = 1000$

 ii) Alternative Hypothesis: $H_1 : \mu < 1000$(Left one tailed test)

 (Since it is given below standard)

 iii) Level of significance : $\alpha = 0.05$

 t tabulated value with 25 degrees of freedom for left tailed test is 1.708

 iv) Test Statistic : $t_{cal} = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n-1}}} = \frac{990 - 1000}{\frac{20}{\sqrt{25}}} = -2.5$

 v) Conclusion: Since $|t_{cal}|$ value $> t_\alpha$ value , we reject H_0

 Hence we conclude that the sample is not upto the standard.

2. **A random sample of size 16 values from a normal population showed a mean of 53 and sum of squares of deviations from the mean equals to 150 . Can this sample be regarded as taken from the population having 56 as mean ? Obtain 95% confidence limits of the mean of the population.?**

Sol: a) Given n = 16

$\bar{x} = 53$

$\mu = 56$ and $\sum(x_i - \bar{x})^2 = 150$

$\therefore S^2 = \frac{\sum(x_i - \bar{x})^2}{n-1} = \frac{150}{15} = 10 \Rightarrow S = \sqrt{10}$

Degrees of freedom ϑ = n-1 = 16-1 =15

i)   Null Hypothesis $H_0 : \mu = 56$

ii)   Alternative Hypothesis $H_1 : \mu \neq 56$ (Two tailed test)

iii)   Level of significance : $\alpha = 0.05$

       t tabulated value with 15 degrees of freedom for two tailed test is 2.13

iv)   Test Statistic : $t_{cal} = \frac{\bar{x} - \mu}{\frac{S}{\sqrt{n}}} = \frac{53 - 56}{\frac{\sqrt{10}}{\sqrt{15}}} = -3.79$

v)   Conclusion: Since $|t_{cal}|\ value > t_\alpha$ value , we reject H_0

       Hence we conclude that the sample cannot be regarded as taken from population.

       b) The 95% confidence limits of the mean of the population are given by

$\bar{x} \pm t_{0.05} \frac{S}{\sqrt{n}} = 53 \pm 2.13 \times 0.79$

$= 53 \pm 1.6827$

$= 54.68$ and 51.31

$\therefore$ 95% confidence limits are(51.31, 54.68)

3. **A random sample of 10 boys had the following I.Q's : 70, 120 ,110, 101,88, 83,95,98,107 and 100.**
   a) **Do these data support the assumption of a population mean I.Q of 100?**
   b) **Find a reasonable range in which most of the mean I.Q values of samples of 10 boys lie**

    Sol: Since mean and s.d are not given

    We have to determine these

x	$x - \bar{x}$	$(x - \bar{x})^2$
70	-27.2	739.84
120	22.8	519.84
110	12.8	163.84
101	3.8	14.44
88	-9.2	84.64
83	-14.2	201.64
95	-2.2	4.84
98	0.8	0.64
107	9.8	96.04
100	2.8	7.84
$\sum x = 972$		$\sum (x - \bar{x})^2 = 1833.60$

Mean , $\bar{x} = \frac{\Sigma x}{n} = \frac{972}{10} = 97.2$ and

$$S^2 = \frac{1}{n-1} \Sigma(x - \bar{x})^2 = \frac{1833.6}{9}$$

$$\therefore S = \sqrt{203.73} = 14.27$$

i)   Null Hypothesis $H_0 : \mu = 100$

ii)   Alternative Hypothesis $H_1 : \mu \neq 100$ (Two tailed test)

iii) Level of significance : $\alpha = 0.05$

    t tabulated value with 9 degrees of freedom for two tailed test is 2.26

iv)   Test Statistic : $t_{cal} = \frac{\bar{x} - \mu}{\frac{S}{\sqrt{n}}} = \frac{97.2 - 100}{\frac{14.27}{\sqrt{10}}} = -0.62$

v)   Conclusion: Since $|t_{cal}|\, value < t_\alpha$ value , we accept H_0

    Hence we conclude that the data support the assumption of mean I.Q of 100 in the population.

    b) The 95% confidence limits of the mean of the population are given by

$$\bar{x} \pm t_{0.05} \frac{S}{\sqrt{n}} = 97.2 \pm 2.26 \times 4.512$$

$$= 97.2 \pm 10.198$$

$$= 107.4 \text{ and } 87$$

$\therefore$ 95% confidence limits are(87, 107.4)

4. **Samples of two types of electric bulbs were tested for length of life and following data were obtained**

Type 1	Type 2
Sample number , $n_1 = 8$	$n_2 = 7$
Sample mean , $\overline{x_1} = 1234$	$\overline{x_2} = 1036$
Sample S.D , $s_1 = 36$	$s_2 = 40$

Is the difference in the mean sufficient to warrant that type 1 is superior to type 2 regarding length of life .

Sol: i) Null Hypothesis H_0 : The two types of electric bulbs are identical

i.e., $H_0: \mu_1 = \mu_2$

ii) Alternative Hypothesis $H_1 : \mu_1 \neq \mu_2$

iii)Test Statistic : $t_{cal} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$

    Where $S^2 = \frac{n_1 s_1^2 + n_1 s_1^2}{n_1 + n_2}$

$$= \frac{1}{8+7-2}(8(36)^2 + 7(40)^2) = 1659.08$$

$\therefore$ t $= \frac{1234 - 1036}{\sqrt{1659.08\left(\frac{1}{8} + \frac{1}{7}\right)}} = 9.39$

iv)Degrees of freedom = 8+7-2 =13 ,tabulated value of t for 13 d.f at 5% los is 2.16

v)Conclusion: Since $|t_{cal}|\, value > t_\alpha$ value , we reject H_0

Hence we conclude that the two types 1 and 2 of electric bulbs are not identical .

5. Two horses A and B were tested according to the time to run a particular track with the following results .

Horse A	28	30	32	33	33	29	34
Horse B	29	30	30	24	27	29	

Test whether the two horses have the same running capacity

Sol: Given $n_1 = 7$, $n_2 = 6$

We first compute the sample means and standard deviations

$\bar{x}$ = Mean of the first sample = $\frac{1}{7}$ (28 + 30 + 32 + 33 + 33 + 29 + 34)

$= \frac{1}{7}(219) = 31.286$

$\bar{y}$ = Mean of the second sample = $\frac{1}{6}$ (29 + 30 + 30 + 24 + 27 + 29)

$= \frac{1}{6}(169) = 28.16$

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	-416	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			
$\sum x$ $= 219$		$\sum (x - \bar{x})^2$ $= 31.4358$	$\sum y$ $= 169$		$\sum (y - \bar{y})^2$ $= 26.8336$

Now $S^2 = \frac{1}{n_1 + n_2 - 2} [(\sum(x - \bar{x})^2 + \sum(y - y)^2]$

$= \frac{1}{11}[31.4358 + 26.8336]$

$= \frac{1}{11}(58.2694)$

$= 5.23$

$\therefore S = \sqrt{5.23} = 2.3$

i)   Null Hypothesis $H_0 : \mu_1 = \mu_2$

ii)   Alternative Hypothesis $H_1 : \mu_1 \neq \mu_2$

iii)   Test Statistic : $t_{cal} = \frac{\bar{x}_1 - \bar{x}_2}{S\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$

$$= \frac{31.286 - 28.16}{2.3\left(\sqrt{\frac{1}{7} + \frac{1}{6}}\right)} = 2.443$$

$\therefore t_{cal} = 2.443$

iv) Degrees of freedom = 7+6-2 =11

Tabulated value of t for 11 d.f at 5% los is 2.2

Conclusion: Since $|t_{cal}|$ value $> t_\alpha$ value , we reject H_0

Hence we conclude that both horses do not have the same running capacity.

6. **Ten soldiers participated in a shooting competition in the first week. After intensive training they participated in the competition in the second week . Their scores before and after training are given below :**

Scores before	67	24	57	55	63	54	56	68	33	43
Scores after	70	38	58	58	56	67	68	75	42	38

Do the data indicate that the soldiers have been benefited by the training.

Sol: Given $n_1 = 10$, $n_2 = 10$

We first compute the sample means and standard deviations

$\bar{x}$ = Mean of the first sample $= \frac{1}{10}$ (67 + 24 + 57 +55+63+54+56+68+33+43)

$= \frac{1}{10}(520) = 52$

$\bar{y}$ = Mean of the second sample $= \frac{1}{10}$ (70+38+58+58+56+67+68+75+42+38)

$= \frac{1}{10}(570) = 57$

x	$x - \bar{x}$	$(x - \bar{x})^2$	y	$y - \bar{y}$	$(y - \bar{y})^2$
67	15	225	70	13	169
24	-28	784	38	-19	361
57	5	25	58	1	1
55	3	9	58	1	1
63	11	121	56	-1	1
54	2	4	67	10	100
56	4	16	68	11	121
68	16	256	75	18	324
33	-19	361	42	-15	225
43	-9	81	38	-19	361
$\sum x = 520$		$\sum (x - \bar{x})^2$ $= 1882$	$\sum y = 570$		$\sum (y - \bar{y})^2$ $= 1664$

Now $S^2 = \frac{1}{n_1+n_2-2}[(\sum(x - \bar{x})^2 + \sum(y - y)^2]$

$= \frac{1}{18}[1882 + 1664]$

$= \frac{1}{18}(3546)$

$= 197$

$\therefore S = \sqrt{197} = 14.0357$

    i)   Null Hypothesis $H_0: \mu_1 = \mu_2$

    ii)   Alternative Hypothesis $H_1 : \mu_1 < \mu_2$ (Left one tailed test)

    iii)   Test Statistic : $t_{cal} = \dfrac{\bar{x}_1 - \bar{x}_2}{S\sqrt{\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$

$$= \frac{52 - 57}{14.0357 \left(\sqrt{\frac{1}{10} + \frac{1}{10}} \right)}$$

$= \frac{3546}{18} = -0.796$

$\therefore t_{cal} = $ -0.796

iv)Degrees of freedom = 10+10-2 =18

Tabulated value of t for 18 d.f at 5% los is -1.734

Conclusion: Since $|t_{cal}|$ value $< |t_\alpha|$ value , we accept H_0

Hence we conclude that the soldiers are not benefited by the training.

7. **The blood pressure of 5 women before and after intake of a certain drug 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% los?

Sol: Given n = 5

i)   Null Hypothesis H_0: $\mu_1 = \mu_2$

ii)   Alternative Hypothesis $H_1 : \mu_1 < \mu_2$ (Left one tailed test)

iii)   Test Statistic $t_{cal} = \frac{\bar{d}}{s/\sqrt{n}}$

where $\bar{d} = \frac{\Sigma d}{n}$ and $S^2 = \frac{1}{n-1} \Sigma (d - \bar{d})^2$

B.P before training	B.P after training	$d = y - x$	$d - \bar{d}$	$(d - \bar{d})^2$
110	120	10	8	64
120	118	-2	-4	16
123	125	2	0	0
132	136	4	2	4
125	121	-4	-6	36
		$\sum d = 10$		$\Sigma (d - \bar{d})^2 = 120$

$\therefore \bar{d} = \frac{10}{5} = 2$ and $S^2 = \frac{120}{4} = 30$

$\therefore S = 5.477$

$t_{cal} = \frac{\bar{d}}{s/\sqrt{n}} = \frac{2}{5.477/\sqrt{5}} = 0.862$

iv)   Degrees of freedom = 5-1= 4

Tabulated value of t for 4 d.f at 1% los is 4.6

Conclusion: Since $|t_{cal}|$ value $< |t_\alpha|$ value , we accept H_0

Hence we conclude that there is no significant difference in Blood pressure after intake of a certain drug.

8. Memory capacity of 10 students were tested before and after training . State whether the training was effective or not from the following scores.

Sol : i) Null Hypothesis H_0: $\mu_1 = \mu_2$

ii)   Alternative Hypothesis $H_1 : \mu_1 < \mu_2$ (Left one tailed test)

iii)   Test Statistic $t_{cal} = \dfrac{\bar{d}}{s/\sqrt{n}}$

where $\bar{d} = \dfrac{\Sigma d}{n}$ and $S^2 = \dfrac{1}{n-1}\Sigma(d - \bar{d})^2$

Before(x)	After(y)	$d = y - x$	d^2
12	15	-3	9
14	16	-2	4
11	10	1	1
8	7	1	1
7	5	2	4
10	12	-2	4
3	10	-7	49
0	2	-2	4
5	3	2	4
6	8	-2	4
		$\sum d = -12$	$\sum d^2 = 84$

$\bar{d} = \dfrac{-12}{10} = $ -1.2

$S^2 = \dfrac{84 - (-1.2)^2 \times 10}{9} = 7.73$

$\therefore$ S = 2.78

$t_{cal} = \dfrac{\bar{d}}{s/\sqrt{n}} = \dfrac{-1.2}{2.78/\sqrt{10}} = $ -1.365 and d.f = n-1 = 9

Tabulated value of t for 9 d.f at 5% los is 1.833

Conclusion: Since $|t_{cal}|$ value $<$ $|t_\alpha|$ value , we accept H_0

Hence we conclude that there is no significant difference in memory capacity after the training program.

Chi-Square (χ^2) Distribution

Chi square distribution is a type of cumulative probability distribution . probability distributions provide the probability of every possible value that may occur . Distributions that are cumulative give the probability of a random variable being less than or equal to a particular value. Since the sum of the probabilities of every possible value must equal one , the total area under the curve is equal to one . Chi square distributions vary depending on the degrees of freedom. The degrees of freedom is found by subtracting one from the number of categories in the data .

.

χ^2
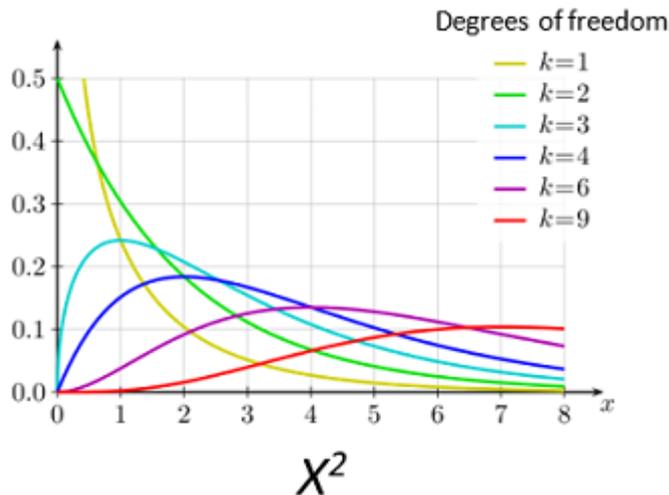
Applications of Chi – Square Distribution:

Chi – Square test as a test of goodness of fit :

χ^2- test enables us to ascertain how well the theoretical distributions such as binomial, Poisson, normal etc, fit the distributions obtained from sample data. If the calculated value of χ^2 is less than the table value at a specified level of generally 5% significance, the fit is considered to be good.

If the calculated value of χ^2 is greater than the table value, the fit is considered to be poor.

i) Null hypothesis: H_0 : There is no difference in given values and calculated values

ii)Altenative hypothesis: H_1 : There is some difference in given values and calculated values

iii) Test Statistic $\chi^2_{cal} = \Sigma \frac{(O-E)^2}{E}$

iv)At specified level of significance for n-1 d.f if the given problem is binomial distribution

At specified level of significance for n-2 d.f if the given problem is Poisson distribution

v)Conclusion :If χ^2_{cal} value $< \chi^2_{tab}$ value , then we accept H_0 , Otherwise reject H_0.

2. Chi – Square test for independence of attributes :

Definition : An attribute means a quality or characteristic
Eg: Drinking, Smoking, blindness, Honesty, beauty etc.,

An attribute may be marked by its presence or absence in a number of a given population.

Let us consider two attributes A and B.

A is divided into two classes and B is divided into two classes. The various cell frequencies can be expressed in the following table known as 2x2 contingency table.

a	b	a + b
c	d	c+ d
a + c	b + d	N =a + b + c + d

The expected frequencies are given by

$$E(a) = \frac{(a + c)(a + b)}{N}$$

$$E(b) = \frac{(b + c)(a + b)}{N}$$

$$E(c) = \frac{(a + c)(c + d)}{N}$$

$$E(d) = \frac{(b + d)(c + d)}{N}$$

$$\chi^2{}_{cal} = \sum \frac{(O - E)^2}{E}$$

$\chi^2{}_{cal}$ value to be compared with $\chi^2{}_{tab}$ value at 1% (5.1 or10%) level of significance for

(r-1) (c-1) d.f where r- number of rows

c-number of columns.

Note: In χ^2 distribution for independence of attributes, we test if two attributes A and B are independent or not.

i)Null Hypothesis: H_0 : The two attributes are independent

ii) Alternative hypothesis: H_1 : The two attributes are not independent

iii) Test Statistic $\chi^2{}_{cal} = \sum \frac{(O - E)^2}{E}$

where $E = \frac{\text{Row total } x \text{ Column total}}{\text{Grand total}}$

iv)At specified level of significance for (m-1) (n-1) d.f where m- no. of rows and n- no. of columns

v)Conclusion : If $\chi^2{}_{cal}$ value $< \chi^2{}_{tab}$ value , then we accept H_0 , Otherwise reject H_0.

Problems :

1. Fit a Poisson distribution to the following data and test for its goodness of fit at 5% los

x	0	1	2	3	4
f	419	352	154	56	19

Sol:

X	f	fx
0	419	0
1	352	352
2	154	308
3	56	168
4	19	76
	N=1000	$\sum fx = 904$

Mean $\lambda = \frac{\sum fx}{N} = \frac{904}{1000} = 0.904$

Theoretical distribution is given by

$= N \times p(x) = 1000 \times \frac{e^{-\lambda}\lambda^x}{x!}$

Hence the theoretical frequencies are given by

x	0	1	2	3	4	Total
f = 1000 x $\frac{e^{-\lambda}\lambda^x}{x!}$	406.2	366	165.4	49.8	12.6	1000

Since Given frequencies total is equal to Calculated frequencies total.

To test for goodness of fit:

i) H_0 : There is no difference in given values and calculated values

ii) H_1 : There is some difference in given values and calculated values

iii) $\chi^2_{cal} = \sum \frac{(O-E)^2}{E}$

O	E	$(O - E)^2$	$\frac{(O - E)^2}{E}$
419	406.2	$(419 - 406.2)^2$	$\frac{(419 - 406.2)^2}{406.2}$
352	366	$(352 - 366)^2$	$\frac{(352 - 366)^2}{366}$
154	165.4	$(154 - 165.4)^2$	$\frac{(154 - 165.4)^2}{165.4}$

56	49.8	$(56 - 49.8)^2$	$\dfrac{(56 - 49.8)^2}{49.8}$
19	12.6	$(19 - 12.6)^2$	$\dfrac{(19 - 12.6)^2}{12.6}$

$$\Sigma \frac{(O-E)^2}{E} = 5.748$$

Degrees of freedom = 5-2 = 3

$\chi^2{}_{tab}$ at 5% LOS = 7.82

Since $\chi^2{}_{cal}$ value $< \chi^2{}_{tab}$,we accept H_0 .

3. **A die is thrown 264 times with following results. Show that the die is biased [Given $\chi^2{}_{0.05} = 11.07$ for 5 d.f]**

No. appeared on the die	1	2	3	4	5	6
Frequency	40	32	28	58	54	52

Sol: i) H_0 : The die is unbiased

ii) H_1 : The die is not unbiased

iii)$\chi^2{}_{cal} = \Sigma \frac{(O-E)^2}{E}$

The expected frequency of each of the number 1,2,3,4,5,6 is $\frac{264}{6} = 44$

Calculation of χ^2:

O	E	$(O - E)^2$	$\dfrac{(O - E)^2}{E}$
40	44	16	0.3636
32	44	144	3.2727
28	44	256	5.8181
58	44	196	4.4545
54	44	100	2.2727
52	44	64	1.4545

$$\Sigma \frac{(O-E)^2}{E} = 17.6362$$

$\chi^2{}_{cal} = 17.6362$

The number of degrees of freedom = n-1 = 5

$\chi^2{}_{0.05} = 11.07$ for 5 d.f

Since $\chi^2{}_{cal}$ value $> \chi^2{}_{tab}$ value , we reject H_0

Hence the die is biased

4. **On the basis of information given below about the treatment of 200 patients suffering from disease , state whether the new treatment is comparatively Superior to the conventional treatment.**

Treatment	Favorable	Not Favorable	Total
New	60	30	90
Conventional	40	70	110

Sol: i) H_0 : The two attributes are independent

ii) H_1 : The two attributes are not independent

iii) $\chi^2_{cal} = \sum \dfrac{(O - E)^2}{E}$

$\dfrac{90 \times 100}{200} = 45$	$\dfrac{90 \times 100}{200} = 45$	90
$\dfrac{100 \times 110}{200} = 55$	$\dfrac{100 \times 110}{200} = 55$	11
100	100	200

where $E = \dfrac{\text{Row total x Column total}}{\text{Grand total}}$

Calculation of χ^2:

O	E	$(O - E)^2$	$\dfrac{(O - E)^2}{E}$
60	45	225	5
30	45	225	5
40	55	225	4.09
70	55	225	4.09

$$\sum \frac{(O-E)^2}{E} = 18.18$$

$\chi^2_{cal} = 18.18$

χ^2_{tab} for 1 d.f . at 5% los is 3.841

since χ^2_{cal} value $> \chi^2_{tab}$ value , we reject H_0

Hence we conclude that new and conventional treatment are not independent.

Snedecor's F- Test of Significance

The F-Distribution is also called as Variance Ratio Distribution as it usually defines the ratio of the variances of the two normally distributed populations. The F-distribution got its name after the name of R.A. Fisher, who studied this test for the first time in 1924.

Symbolically, the quantity is distributed as F-distribution with and degrees of freedom $\vartheta_1 = n_1 - 1$ and $\vartheta_2 = n_2 - 1$ is represented as:

$$F_{cal} = \frac{\text{Greater Variance}}{\text{Smaller Varinace}}$$

$$F_{cal} = \frac{S_1{}^2}{S_2{}^2} \text{ Or } \frac{S_2{}^2}{S_1{}^2}$$

Where,

$S_1{}^2$ is the unbiased estimator of σ_1^2 and is calculated as: $S_1{}^2 = \frac{n_1 s_1{}^2}{n_1 - 1} = \frac{1}{n_1 - 1} \sum (x_1 - \overline{x_1})^2$

$S_2{}^2$ is the unbiased estimator of σ_2^2 and is calculated as: $S_2{}^2 = \frac{n_2 s_2{}^2}{n_2 - 1} = \frac{1}{n_2 - 1} \sum (x_2 - \overline{x_2})^2$

To test the hypothesis that the two population variances $\sigma_1{}^2$ and $\sigma_2{}^2$ are equal

i) $H_0 : \sigma_1{}^2 = \sigma_2{}^2$

ii) $H_1 : \sigma_1{}^2 \neq \sigma_2{}^2$

iii) $F_{cal} = \frac{\text{Greater Variance}}{\text{Smaller Varinace}}$

iv) At specified level of significance (1% or 5 %) for $(\vartheta_1, \vartheta_2)$ d.f

v) If F_{cal} value $< F_{tab}$ value , then we accept H_0 , Otherwise reject H_0.

$F_{cal}(\vartheta_1, \vartheta_2)$ is the value of F with ϑ_1 and ϑ_2 degrees of freedom such that the area under the F – distribution to the right of F_α is α.
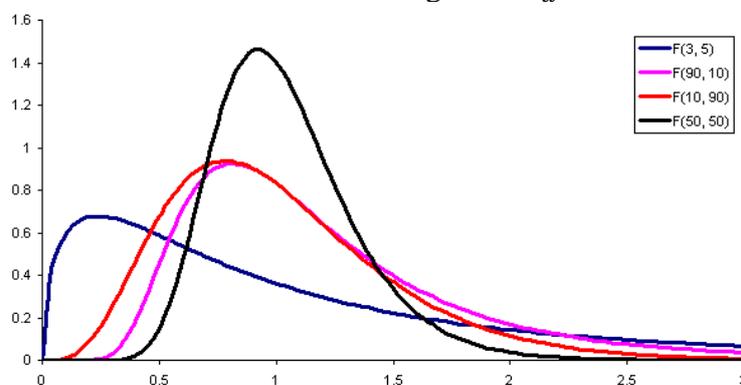
Problems:

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 varience.**

Sol: Let σ_1^2 and σ_2^2 be the variances of the two normal populations from which the samples are drawn.

Let the Null Hypothesis be $H_0: \sigma_1^2 = \sigma_2^2$

Then the Alternative Hypothesis is $H_1: \sigma_1^2 \neq \sigma_2^2$

Here $n_1 = 8, n_2 = 10$

Also $\Sigma(x_i - \bar{x})^2 = 84.4, \Sigma(y_i - \bar{y})^2 = 102.6$

If $S_1^2 and S_2^2$ be the estimates of σ_1^2 and σ_2^2 then

$$S_1^2 = \frac{1}{n_1 - 1}\sum(x_i - \bar{x})^2 = \frac{84.4}{7} = 12.057$$

and

$$S_2^2 = \frac{1}{n_2 - 1}\sum(y_i - \bar{y})^2 = \frac{102.6}{9} = 11.4$$

Let H_0 be true. Since $S_1^2 > S_2^2$, the test statistic is

$F = \frac{S_1^2}{S_2^2} = \frac{12.057}{11.4} = 1.057$

i.e., calculated F = 1.057.

Degrees of freedom are given by $v_1 = n_1 - 1 = 8 - 1 = 7$

and $v_2 = n_2 - 1 = 10 - 1 = 9$

Tabulated value of F at 5% level for (7,9) degrees of freedom is 3.29

i.e., $F_{0.05}(7,9) = 3.29$

Since calculated $F <$ tabulated F, we accept the Null Hypothesis H_0 and conclude that the populations have the same variance.

2.The time taken by workers in performing a job by method I and method II is given below

Method I	20	16	26	27	23	22	-
Method II	27	33	42	35	32	34	38

Do the data show that the variances of time distribution from population from which these samples are drawn do not differ significantly?

Sol: Let the Null Hypothesis be $H_0: \sigma_1{}^2 = \sigma_2{}^2$ where $\sigma_1{}^2$ and $\sigma_2{}^2$ are the variances of the two populations from with the samples are drawn.

The Alternative Hypothesis is $H_1: \sigma_1{}^2 \neq \sigma_2{}^2$.

Calculation of sample variances.

x	$x - \bar{x}$	$(x - \bar{x})^2$	y	$y - \bar{y}$	$(y - \bar{y})^2$
20	-2.3	5.29	27	-7.4	54.76
16	-6.3	39.69	33	-1.4	1.96
26	3.7	13.69	42	7.6	57.76
27	4.7	22.09	35	0.6	0.36
23	0.7	0.49	32	-2.4	5.76
22	-0.3	0.09	34	-0.4	0.16
			38	3.6	12.96
134		81.34	241		133.72

$n_1 = 6, n_2 = 7$

$$\therefore \bar{x} = \frac{\sum x}{n_1} = \frac{134}{6} = 22.3, \bar{y} = \frac{\sum y}{n_2} = \frac{241}{67} = 34.$$

$$\sum (x_i - \bar{x})^2 = 81.34, \sum (y_i - \bar{y})^2 = 133.72$$

If $S_1{}^2 and S_2{}^2$ be the estimates of $\sigma_1{}^2$ and $\sigma_2{}^2$, then

$$S_1{}^2 = \frac{1}{n_1 - 1}\sum (x_i - \bar{x})^2 = \frac{81.34}{5} = 16.26$$

and

$$S_2{}^2 = \frac{1}{n_2 - 1}\sum (y_i - \bar{y})^2 = \frac{133.72}{6} = 22.29$$

Let H_0 be true

Since $S_2{}^2 > S_1{}^2$, the statistic is

$$F = \frac{S_2{}^2}{S_1{}^2} = \frac{22.29}{16.268} = 1.3699 = 1.37$$

$F_{0.05}(5,6) \, d.f = 4.39$

Since calculated $F <$ tabulated F, we accept the null hypothesis H_0 at 5% los i.e., there is no significant difference between the variances of the distribution by the workers.

TUTORIAL QUESTIONS

1. A random sample of 500 apples was taken from a large consignment and 60 were found to be bad. Obtain 95% confidence interval for the percentage number of bad apples in the consignment.

2. The average income of 100 people of a city is Rs 210 with a standard deviation of Rs 10.For another sample of 150 people the average income is Rs 220 with a standard deviation of Rs 12.Test the significant difference between two mean at 5% LOS.

3. A coin is tossed 960 times .Head turned up 184 times. Find whether the coin is unbiased.

4. Random samples of 600 men and 900 women in a locality were asked they would like to have a bus stop near their residence .350 men and 475 women were in favor of the proposal. Test the significance between the difference of two proportions at 5%LOS.

5. .A pair of dice are thrown 360 times and the frequency of each sum is indicated below:

Sum	2	3	4	5	6	7	8	9	10	11	12
Frequency	8	24	35	37	44	65	51	42	26	14	14

Would you say that the dice are fair on the basis of the chi-square test at 5% LOS

6. The following are the average weekly losses of worker hours due to accidents in 10 industrial plant before and after a certain safety programme was put into operation:

Before	45	73	46	124	33	57	83	34	26	17
After	36	60	44	119	35	51	77	29	24	11

Test whether the safety programme is effective in reducing the number of accidents at 5%

ASSIGNMENT QUESTIONS

1. A random sample of 500 items has mean 20 and another sample of size 400 has mean 15.

   Can you conclude that the two samples are taken from the same population with SD as 4.

2. A sample of 500 products are examined from a factory and 5% found to be defective. Another sample of 400 similar products are examined and 3% found to be defective. Test the significance between the difference of two proportions at 5% LOS.

3. 20 people were attacked by a disease and only 18 survived .will you reject the hypothesis that the survival rate of the attack by this disease is 85% in favor of the hypothesis that is more at 5% LOS

4. Ten specimens of copper wires drawn from a large lot have the following breaking strength(in kg) 518,572,570,568,572,578,572,569,548.Test whether the mean breaking strengths of the lot may be taken to be 518 kg weight.

5. 4.A survey of 320 families with 4children each revealed the following distribution

No# of boys	5	4	3	2	1	0
No# of girls	0	1	2	3	4	5
No# of families	14	56	110	88	40	12

   Is this result consistent with the hypothesis that male and female births are equally

   popular?

R15

MALLA REDDY COLLEGE OF ENGINEERING & TECHNOLOGY
(Autonomous Institution – UGC, Govt. of India)
II B.Tech I Semester Regular/Supplementary Examinations, November 2017
Probability and Statistics
(CSE, IT)

Roll No										

Time: 3 hours   **Max. Marks: 75**

Note: This question paper contains two parts A and B

Part A is compulsory which carriers 25 marks and Answer all questions.

Part B Consists of 5 SECTIONS (One SECTION for each UNIT). Answer FIVE Questions, Choosing ONE Question from each SECTION and each Question carries 10 marks.

PART – A   (25 Marks)

1. (a) What are the measures of Central tendency?

   (b) A box contains 6 red, 4 white and 5 black balls. A person draws 4 balls from the box at random. Find the probability that among the balls drawn there is at least one ball of each color.

   (c) What is principle of least square ?

   (d) In a record of an analysis of correlation data, the following results are readable: Variance of X = 9; Regression equations: 8X-10Y+66 = 0 and 40X-18Y = 214. Find the mean values of X and Y.

   (e) Define sampling distribution.

   (f) What are the types of errors in testing of hypothesis?

   (g) State applications of student t- distribution.

   (h) Write all the properties of Chi-square distribution.

   (i) What are the characteristics of a queue in queuing system?

   (j) Write a short note on stochastic process.

PART – B   (50 Marks)

SECTION – I

2. a) State and prove addition law of probability for two events

   b) In a bolt factory machines A_1, A_2, A_3 manufacture respectively 25%, 35% and 40% of the total output. Of these 5, 4, and 2 percent are defective bolts. A bolt is drawn at random from the product and is found to be defective. What is the probability that it was manufactured by machine A_2.

(OR)

3. Find the mean and variance of binomial distribution.

SECTION – II

4. Calculate the correlation coefficient for the following heights (in inches) of fathers (X) and their sons (Y):

X : 65   66   67   67   68   69   70   72

Y : 67   68   65   68   72   72   69   71

(OR)

5. Fit a linear regression equation of Y on X to the following data:

X: 5   8   7   6   4

Y: 3   4   5   2   1

SECTION – III

6. A sample of 900 members has a mean 3.4 cms and s.d 2.61 cms. Is the sample drawn from a large population of mean 3.25 cms and s.d 2.61 cms? 5% level.

(OR)

7. In a sample of 1000 people in a state, 540 are rice eaters and the rest are wheat eaters. Can we assume that both rice and wheat eaters are equally popular in this state at 1% level of significance?

SECTION – IV

8. A sample of 10 boys has the I.Q's 70, 120, 110,101, 88, 83, 95, 98, 107 and 100. Test the mean I.Q of the students is 100 at 0.05 level of significance.

(OR)

9. A survey of 320 families with 5 children each, revealed the following distribution. Is the result consistent with the hypothesis that male and female births are equally probable at 0.01 significance level?

No. of Boys	:	5	4	3	2	1	0
No. of Girls	:	0	1	2	3	4	5
No. of families:		14	6	110	88	40	12

SECTION – V

10. A television repairman finds that the time spent on his jobs an exponential distribution with mean 30 minutes. If he repairs sets in the order in which they came in, and if the arrival of sets follows a poisson distribution approximately with an average rate of 10 per 8-hour daily, what is the repairman's expected idle time each day? How many jobs are a head of the average set just brought in?

(OR)

11. Explain Markov chain by an example.

******

R17

MALLA REDDY COLLEGE OF ENGINEERING & TECHNOLOGY
(Autonomous Institution – UGC, Govt. of India)
II B. Tech I Semester Regular Examinations, November 2018
Probability and Statistics
(CSE& IT)

Roll No									

Time: 3 hours   **Max. Marks: 70**

Note: This question paper Consists of 5 Sections. Answer **FIVE** Questions, Choosing ONE Question from each SECTION and each Question carries 14 marks.

SECTION-I

1   a) Average number of accidents on any day on a national highway is 1.8. Determine the probability that number of accidents are   **[7M]**
          (I)   At least one
          (II)   Atthe most one

      b) Four coins are tossed 160 times. The number of times X heads occurs is given below   **[7M]**

X	0	1	2	3	4
No of times	8	34	69	43	6

          Fit the binomial distribution for the above data

OR

2   a) A random variable X has the following probability distribution   **[7M]**

                          X   -3   6   9
                          P(X=x) 1/6   1/2   1/3

          Find (1) Mean and variance (2) Find E[Y], Var[Y], given that Y=2X+1.   **[7M]**

      b) A r.v X is a normally distributed with mean 30 and SD 5.Find the probabilities that (i)$26 \leq X \leq 40$ (ii) $X \geq 45$

SECTION-II

3   a) Fit a linear regression equation of Y on X to the following data: **[7M]**

X:	5	8	7	6	4
Y:	3	4	5		1

      b) Find the spearman rank correlation coefficient to the following data: **[7M]**
            X: 11 12   43   84   15
            Y: 8   15   30   60   12

OR

4   a)In a record of an analysis of correlation data, the following results are readable: Variance of X = 9   **[7M]**
      Regression equations: 8X-10Y+66 = 0 and
      40X-18Y = 214. Find (i) the mean values of X and Y (ii) The correlation coefficient between X and Y
      and (iii) The standard deviation of Y

      b) Calculate the correlation coefficient for the following heights (in inches) of fathers (X) and their   **[7M]**
      sons (Y):
      X : 65   66   67   67   68   69   70 72
      Y : 67   68   65   68   72   72   69 71

SECTION-III

5   a) Explain in brief one tailed and two tailed tests   **[4M]**

      b) A random sample of 400 students is found to have a mean height of 171.38 cms. Can it be reasonably regarded as a sample from a large population with mean height 171.17 cms. and standard deviation 3.30 cms. (Test at 5% level of significance   **[5M]**

      c) A random sample of 500 apples was taken from a large consignment and 60 were found

bad. Obtain the 98% confidence limits for the percentage of bad apples in the consignment (given z = 2.33) **[5M]**

OR

6 Random samples of 400 men and 600 women were asked whether they would like to have a flyover **[14M]** 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 same at 5% level.

SECTION-IV

7 A sample of 10 boys has the I.Q's 70, 120, 110,101, 88, 83, 95, 98, 107 and 100. Test the mean I.Q **[14M]** of the students is 100 at 0.05 level of significance.

OR

8 Fit a Poisson distribution to the following data and test the goodness of fit: **[14M]**

No. of accidents:	0	1	2	3	4	5	6
No. of days :	150	65	45	34	10	6	2

SECTION-V

9 What are the measures of queuing model $M/M/1: N/FCFS$. **[14M]**

b A self service canteen employs one cashier at its counter. 8 customers arrive per every 10 minutes on an average. The cashier can serve on average one per minute. Assuming that arrivals are Poisson and the service time distribution is exponential. Determine

i) The average number of customers in the system
ii) The average queue length
iii) The average time a customer spends in the system
iv) Average waiting time of each customer.

OR

10 a)Define Markov chain.Give examples. **[14M]**
b) Explai about limiting distribution of a Markov chain

****

R15

MALLA REDDY COLLEGE OF ENGINEERING & TECHNOLOGY
(Autonomous Institution – UGC, Govt. of India)
II B.Tech I Semester supplementary Examinations, November 2018
Probability and Statistics
(CSE &IT)

Roll No									

Time: 3 hours **Max. Marks: 75**

Note: This question paper contains two parts A and B

Part A is compulsory which carriers 25 marks and Answer all questions.

Part B Consists of 5 SECTIONS (One SECTION for each UNIT). Answer FIVE Questions, Choosing ONE Question from each SECTION and each Question carries 10 marks.

****

PART-A (25 Marks)

1). a If X denote the number of heads in a single toss of 4 fair coins. **[2M]**
Determine P $(1 < X \leq 3)$.

b The mean of a poisson distribution is 3 find $P(X < 2)$. **[3M]**

c Write the equations of lines of regression. **[2M]**

d Show that correlation coefficient is the geometric mean between the two regression coefficients. **[3M]**

e Define Null hypothesis and Alternative hypothesis. **[2M]**

f Write the test of statistic for difference of proportions. **[3M]**

g Write properties of F-distribution. **[2M]**

h Write a short note on degrees of freedom. **[3M]**

i Write a short note on queue behavior. **[2M]**

j Test whether matrix is stochastic or not $\begin{bmatrix} 15/6 & 1/16 \\ 2/3 & 4/3 \end{bmatrix}$ **[3M]**

PART-B (50 MARKS)

SECTION-I

2 Suppose 5 men out of 100 and 25 women out of 10,000 are colour blind. A colour blind person is chosen at random. What is the probability of the person being a male? **[10M]**

OR

3 Fit a Poisson distribution for the following data and calculate the expected frequencies **[10M]**

x	0	1	2	3	4
F(x)	109	65	22	3	1

SECTION-II

4 Find the correlation coefficient between x and y **[10M]**

x	1	2	3	4	5	6	7	8	9
y	12	11	13	15	14	17	16	19	18

OR

5 Test whether the equations $2x + 3y = 4$ and $x - y = 5$ represent valid regression lines. **[10M]**

SECTION-III

6 The average mark scored by 32 boys is 72 with a standard deviation of 8.While that for 36 girls is 70 with a standard deviation of 6. Does this indicate that the boys perform better than girls at a level of significance? **[10M]**

OR

7 In a big city 325men out of 600men were found to be smokers. Does this information support the conclusion that the majority of men in this city are smokers? **[10M]**

SECTION-IV

8 200 digits were chosen at random from a set of tables. The frequencies of the digits were **[10M]**

Digits	0	1	2	3	4	5	6	7	8	9
Frequency	18	19	23	21	16	25	22	20	21	15

Use χ^2 test to assess the correctness of the hypothesis that the digits were distributed in equal numbers in the table at 0.05 level of significance.

OR

9 Two horses A and B were tested according to the time (in seconds) to run a particular track with the following results. **[10M]**

Horse A	28	30	32	33	33	29	34
Horse B	29	30	30	24	27	29	

Test whether the two horses have the same running capacity.

SECTION-V

10 Assume the goods trains are coming in a yard at the rate of 30 trains per day and suppose that inter-arrival time follows an exponential distribution. The service time for each train is assumed to be exponential with an average of 36 minutes. If the yard can admit 9 trains at a time, calculate the probability that the yard is empty and find the average queue length. **[10M]**

OR

11 A training process is considered as two state Markov chain. If it rains, it is considered to be in state 0 and it does not rain, the chain is in state of 1. The transition probability of the Markov chain is defined by P=$\begin{bmatrix} 0.6 & 0.4 \\ 0.2 & 0.8 \end{bmatrix}$. Find the probability that it will rain for three days from today assuming that is raining today. Assume the mutual probabilities of state 0 or state 1 as 0.4 and 0.6 respectively. **[10M]**

******

MALLA REDDY COLLEGE OF ENGINEERING & TECHNOLOGY
(Autonomous Institution – UGC, Govt. of India)
II B.Tech I Semester Supplementary Examinations, May 2019
Probability and Statistics
(CSE & IT)

Roll No									

Time: 3 hours   **Max. Marks: 70**

Note: This question paper Consists of 5 Sections. Answer **FIVE** Questions, Choosing ONE Question from each SECTION and each Question carries 14 marks.

***

SECTION-I

1   a.) A random variable X has the probability function, find k and mean   **[7M]**

X=x	1	2	3	4	5	6
P(X)	K	3k	5k	7k	9k	11k

   b) A manufacturer of television set known that on an average 5% of their   **[7M]**
   product is defective. They sells television sets in consignment of 100 and
   guarantees that not more than 2 set will be defective. What is the probability
   that the TV set will fail to meet the guaranteed quality?

OR

2   A sales tax officer has reported that the average sales of the 500 business that he   **[14M]**
has to deal with during a year is Rs.36,000 with a standard deviation of Rs.10,000.
Assuming that the sales in these business are normally distributed ,find :
i) The number of business as the sales of which are greater than Rs.40,000
ii)The percentage of business the sales of which are likely to range between
Rs.30,000 and Rs.40,000

SECTION-II

3   Find the karlpearson's coefficient of correlation for the paired data: **[14M]**

wages	100	101	102	100	99	97	98	96	95	102
Cost of living	98	99	99	95	92	95	94	90	91	97

OR

4   a)In a record of an analysis of correlation data, the following results are readable   **[7M]**
Variance of X = 9; Regression equations: 8X-10Y+66 = 0 and
40X-18Y = 214. Find (i) the mean values of X and Y (ii) The correlation coefficient
between X and Y and (iii) The standard deviation of Y

   b) Find the spearman rank correlation coefficient to the following data: **[7M]**

X: 11   12   43   84   15
Y: 8   15   30   60   12

SECTION-III

5   a) Explain in brief. (1) Type-I-error (2) Type-II-error   **[3M]**
   b) Discuss the test procedure for testing single mean of the population when   **[4M]**
   size of the sample is large.
   c) Nine Determinations of specific heat of iron had a SD of 0.0086.   **[7M]**

Assuming that the determinatins constitute a random sample from normal population, test the claim that the SD σ is less than 0.01 at 95% confidence.

<div align="center">OR</div>

6   a) Random samples of 400 men and 600 women were asked whether they would like to have a flyover 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 same at 5% level. **[7M]**

       b) In a large consignment of oranges random sample of 64 oranges revealed that14 oranges were band. Is it reasonable to ensure that 20% are bad? **[7M]**

<div align="center">

SECTION-IV

</div>

7   Some investigators have proposed that students have elevated blood pressure during finals week. To test this hypothesis 8 students volunteered to have their blood pressure taken at the beginning of the semester and then again during finals week. The blood pressure data (diastolic) is listed below. Test the hypothesis to determine if there is a significant difference in blood pressure levels between the first week and finals week. **[14M]**

Student:	A	B	C	D	E	F	G	H
1ˢᵗ Week:	89	76	84	89	74	66	56	74
Final:	92	78	92	94	93	74	65	73

                                        (*table value $t_{(0.05,\ 7)} = 2.365$*)

<div align="center">OR</div>

8   Fit a Poisson distribution to the following data and test the goodness of fit: **[14M]**

No. of accidents:	0	1	2	3	4	5	6
No. of days :	150	65	45	34	10	6	2

<div align="center">

SECTION-V

</div>

9   a)What are the measures of queuing model $M/M/1:\infty/FCFS$ **[4M]**

    b) In railway marshalling yard goods trains arrive at a rate of 30 trains per day. Assuming that the inter arrival time follows an exponential distribution and service time distribution is also exponential with an average 36 minutes. Calculate: **[10M]**
    (1) The mean queue size
    (2) The probability that the queue length exceeds 10.

<div align="center">OR</div>

10   A gambler has Rs.2. He bets Rs.1 at a time and wins Rs.1 with probability 0.5. He stops playing if he loses Rs.2 or wins Rs.4. **[14M]**
    a) What is the transition probability matrix of the related Markov chain?
    b) What is the probability that he has lost his money at the end of 5 plays?

<div align="center">**********</div>

MALLA REDDY COLLEGE OF ENGINEERING & TECHNOLOGY
(Autonomous Institution – UGC, Govt. of India)
II B. Tech I Semester Supplementary Examinations, May 2018
Probability and Statics
(CSE & IT)

Roll No									

Time: 3 hours **Max. Marks: 75**

Note: This question paper contains two parts A and B

Part A is compulsory which carriers 25 marks and Answer all questions.

Part B Consists of 5 SECTIONS (One SECTION for each UNIT). Answer FIVE Questions, Choosing ONE Question from each SECTION and each Question carries 10 marks.

***

PART – A (25 Marks)

1. (a) Define Random variable? (2M)

(b) A problem in statistics is given to three students A, B and C, whose chances of solving it are respectively.1/2, 1/3 and 1/4. What is the probability that the problem will be solved? (3M)

(c) What are the normal equations to fit a straight line equation? (2M)

(d) From the following data, compute the coefficient of correlation between X and Y. (3M)

	X Series	Y Series
No. of Items	15	15
Arithmetic Mean	25	18
Sum of squares of deviations from mean:	136	138
Sum of products of deviations of X and Y from their means :	122	

.

(e) Define standard error of an estimate. (2M)

(f) Explain types of errors in tests of significance? (3M)

(g) State applications of Chi-square distribution. (2M)

(h) Write all the properties of t- distribution. (3M)

(i) What is a queue in queuing system? (2M)

(j) Write a short note on Markov chain. (3M)

PART – B (50 Marks)

SECTION – I

2. a) State Baye's theorem. (3M)

b). A random variable X has the following probability function. (7M)

$$\begin{array}{c|cccccccc} X & 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ P(X) & 0 & k & 2k & 2k & 3k & k2 & 2k2 & 7k2+k \end{array}$$

Find (i) k (ii) E(x)

(OR)

3. X is a normally distributed with mean $\mu = 30$ and SD $\sigma = 4$. Find a) $P(x < 40)$ b) $P(x > 21)$ c) $P(30 < x < 35)$.   (10M)

SECTION – II

4. Find the spearman rank correlation coefficient to the following data: (10M)

X:	11	12	43	84	15
Y:	8	15	30	60	12

(OR)

5. Estimate the production for the year 2008, by fitting regression line to the following data: (10M)

Year:	2003	2004	2005	2006	2007
Production: (in thousand Qt).	5	8	14	12	13

.

SECTION – III

6. The means of two large samples of 1000 and 2000 items are 67.5 cms and 68.0cms respectively. Can the samples be regarded as drawn from the population with standard deviation 2.5 cms. Test at 5% level of significance.. (10M)

(OR)

7. A random sample of 500 apples was taken from a large consignment and 60 were found bad. Obtain the 98% confidence limits for the percentage of bad apples in the consignment. (given z = 2.33) (10M)

SECTION – IV

8. Fit a Poisson distribution to the following data and test the goodness of fit: (10M)

No. of accidents :	0	1	2	3	4	5	6
No. of days :	150	65	45	34	10	6	2 .

(OR)

9. A sample of 10 boys has the I.Q's 70, 120, 110,101, 88, 83, 95, 98, 107 and 100. Test the mean I.Q of the students is 100 at 0.05 level of significance. (10M)

SECTION – V

10. What are the characteristics of a queuing system explain them in detail? (10M)

(OR)

11. Explain stochastic processes in detail. (10M)

****