INSTITUTE OF AERONAUTICAL ENGINEERING

(Autonomous)

Dundigal, Hyderabad -500 043

AERONAUTICAL ENGINEERING

COURSE LECTURE NOTES

Course Name	**PROBABILITY AND STATISTICS**
Course Code	AHSB12
Program	B. Tech
Semester	III
Course Coordinator	Mr. CH Chaitanya
Lecture Numbers	1-63
Topic Covered	All

COURSE OBJECTIVES:

The students will try to learn:	
I	The Principles of probability, the theory of random variables, basic random variate distributions and their applications.
II	The Methods and techniques for quantifying the degree of closeness among two or more variables and linear regression analysis.
III	The Estimation statistics and Hypothesis testing which play a vital role in the assessment of the quality of the materials, products and ensuring the standards of the engineering process.
IV	The statistical tools which are essential for translating an engineering problem into probability model.

COURSE OUTCOMES:

After successful completion of the course, students will be able to:

	Course Outcomes	**Knowledge Level (Bloom's Taxonomy)**
CO 1	**Determine** the conditional probability of interdependent events by using Bayes theorem.	Apply
CO 2	**Explain** simulation of random events by using the concept of random variables	Understand
CO 3	**Calculate** the expected values, variances of the discrete and continuous random variables for making decisions under randomized probabilistic conditions.	Apply

CO 4	**Interpret** the Probability distributions such as Binomial, Poisson and Normal distribution by using their probability functions and parameters.	Understand
CO 5	**Apply** the concepts of discrete and continuous probability distribution and CLT for solving real time problems under probabilistic conditions.	Apply
CO 6	**Interpret** the results of Bivariate and Multivariate Regression as well as Correlation Analysis for statistical forecasting.	Understand
CO 7	**Identify** the role of types of statistical hypotheses, types of errors, sampling distributions of means and confidence intervals in hypothesis testing	Apply
CO 8	**Apply** tests of hypotheses for both large and small samples in making decisions over statistical claims.	Apply
CO 9	**Test for** the assessment of goodness of fit of the given probability distribution model by using Chi-square distribution.	Analyze
CO 10	**Make Use of** R software package in computing confidence intervals, Regression analysis and hypothesis testing.	Apply
CO 11	**Select** appropriate statistical methods for solving real-time engineering problems governed by laws of probability.	Apply

SYLLABUS

MODULE-I	**PROBABILITY AND RANDOM VARIABLES**
Probability, Conditional Probability, Baye's Theorem; Random variables: Basic definitions, discrete and continuous random variables; Probability distribution: Probability mass function and probability density functions; Mathematical expectation.	
MODULE-II	**PROBABILITY DISTRIBUTION**
Binomial distribution; Mean and variances of Binomial distribution, Recurrence formula for the Binomial distribution; Poisson distribution: Poisson distribution as a limiting case of Binomial distribution, mean and variance of Poisson distribution, Recurrence formula for the Poisson distribution; Normal distribution; Mean, Variance, Mode, Median, Characteristics of normal distribution.	
MODULE-III	**CORRELATION AND REGRESSION**
Correlation: Karl Pearson's Coefficient of correlation, Computation of correlation coefficient, Rank correlation, Repeated Ranks; Properties of correlation. Regression: Lines of regression, Regression coefficient, Properties of Regression coefficient, Angle between two lines of regression; Multiple correlation and Regression.	
MODULE-IV	**TEST OF HYPOTHESIS - I**
Sampling: Definitions of population, Sampling, Parameter of statistics, standard error; Test of significance: Null hypothesis, alternate hypothesis, type I and type II errors, critical region, confidence interval, level of significance. One sided test, two sided test. Large sample test: Test of significance for single mean, Test of significance for difference between two sample means, Tests of significance single proportion and Test of difference between proportions.	
MODULE-V	**TEST OF HYPOTHESIS - II**
Small sample tests: Student t-distribution, its properties: Test of significance difference between sample mean and population mean; difference between means of two small samples. Snedecor's F-distribution and its properties; Test of equality of two population variances Chi-square distribution and it's properties; Chi-square test of goodness of fit.	

Text Books:

1. Erwin Kreyszig, “Advanced Engineering Mathematics”, John Wiley & Sons Publishers, 9th Edition, 2014.
2. B. S.Grewal, “Higher Engineering Mathematics”, Khanna Publishers, 43rd Edition, 2012.

Reference Books:

1. S. C. Gupta, V. K. Kapoor, “Fundamentals of Mathematical Statistics”, S. Chand & Co., 10th Edition, 2000.
2. N. P. Bali, “Engineering Mathematics”, Laxmi Publications, 9th Edition, 2016.
3. Richard Arnold Johnson, Irwin Miller and John E. Freund, “Probability and Statistics for Engineers”, Prentice Hall, 8thEdition, 2013.

MODULE-I

PROBABILITY AND RANDOM VARIABLES

Probability:

Probability is a branch of mathematics that deals with calculating the likelihood of a given event's occurrence, which is expressed as a number between 1 and 0. An event with a probability of 1 can be considered a certainty: for example, the probability of a coin toss resulting in either "heads" or "tails" is 1, because there are no other options, assuming the coin lands flat. An event with a probability of .5 can be considered to have equal odds of occurring or not occurring: for example, the probability of a coin toss resulting in "heads" is .5, because the toss is equally as likely to result in "tails." An event with a probability of 0 can be considered an impossibility: for example, the probability that the coin will land (flat) without either side facing up is 0, because either "heads" or "tails" must be facing up. A little paradoxical, probability theory applies precise calculations to quantify uncertain measures of random events.

In its simplest form, probability can be expressed mathematically as: the number of occurrences of a targeted event divided by the number of occurrences *plus* the number of failures of occurrences (this adds up to the total of possible outcomes):

$$p(a) = p(a)/[p(a) + p(b)]$$

Calculating probabilities in a situation like a coin toss is straightforward, because the outcomes are mutually exclusive: either one event or the other must occur. Each coin toss is an *independent* event; the outcome of one trial has no effect on subsequent ones. No matter how many consecutive times one side lands facing up, the probability that it will do so at the next toss is always .5 (50-50). The mistaken idea that a number of consecutive results (six "heads" for example) makes it more likely that the next toss will result in a "tails" is known as the *gambler's fallacy* , one that has led to the downfall of many a bettor.

Probability theory had its start in the 17th century, when two French mathematicians, Blaise Pascal and Pierre de Fermat carried on a correspondence discussing mathematical problems dealing with games of chance. Contemporary applications of probability theory run the gamut of human inquiry, and include aspects of computer programming, astrophysics, music, weather prediction, and medicine.

Trial and Event: Consider an experiment, which though repeated under essential and identical conditions, does not give a unique result but may result in any one of the several possible outcomes. The experiment is known as **Trial** and the outcome is called **Event**

E.g. (1) Throwing a dice experiment getting the no’s 1,2,3,4,5,6 (event)

(2) Tossing a coin experiment and getting head or tail (event)

Exhaustive Events:

The total no. of possible outcomes in any trial is called exhaustive event.

E.g.: (1) In tossing of a coin experiment there are two exhaustive events.

(2) In throwing an n-dice experiment, there are 6^n exhaustive events.

Favorable event:

The no of cases favorable to an event in a trial is the no of outcomes which entities the happening of the event.

E.g. (1) In tossing a coin, there is one and only one favorable case to get either head or tail.

Mutually exclusive Event: If two or more of them cannot happen simultaneously in the same trial then the event are called mutually exclusive event.

E.g. In throwing a dice experiment, the events 1,2,3,------6 are M.E. events

Equally likely Events: Outcomes of events are said to be equally likely if there is no reason for one to be preferred over other. E.g. tossing a coin. Chance of getting 1,2,3,4,5,6 is equally likely.

Independent Event:

Several events are said to be independent if the happening or the non-happening of the event is not affected by the concerning of the occurrence of any one of the remaining events.

An event that always happen is called **Certain event,** it is denoted by 'S'.

An event that never happens is called **Impossible event**, it is denoted by 'ϕ'.

Eg: In tossing a coin and throwing a die, getting head or tail is independent of getting no's 1 or 2 or 3 or 4 or 5 or 6.

Definition: probability (Mathematical Definition)

If a trial results in n-exhaustive mutually exclusive, and equally likely cases and m of them are favorable to the happening of an event E then the probability of an event E is denoted by P(E) and is defined as

$$P(E) = \frac{no\ of\ \ favourable\, cases\ to\ event}{Total\ no\ of\ \ exaustive\ cases} = \frac{m}{n}$$

Sample Space:

The set of all possible outcomes of a random experiment is called Sample Space .The elements of this set are called sample points. Sample Space is denoted by S.

Eg. (1) In throwing two dies experiment, Sample S contains 36 Sample points.

S = {(1,1) ,(1,2) ,----------(1,6), --------(6,1),(6,2),--------(6,6)}

Eg. (2) In tossing two coins experiment , S = {HH ,HT,TH,TT}

A sample space is called **discrete** if it contains only finitely or infinitely many points which can be arranged into a simple sequence $w_1, w_2, \ldots\ldots$ .while a sample space containing non denumerable no. of points is called a continuous sample space.

Statistical or Empirical Probability:

If a trial is repeated a no. of times under essential homogenous and identical conditions, then the limiting value of the ratio of the no. of times the event happens to the total no. of trials, as the number of trials become indefinitely large, is called the probability of happening of the event.(It is assumed the limit is finite and unique)

Symbolically, if in 'n' trials and events E happens 'm' times , then the probability 'p' of the happening of E is given by p = P(E) = $\lim\limits_{n\to\infty} \frac{m}{n}$.

An event E is called **elementary event** if it consists only one element.

An event, which is not elementary, is called **compound event.**

Example 1: What is the probability of getting a 2 or a 5 when a die is rolled?

Solution:

Taking the individual probabilities of each number, getting a 2 is 1/6 and so is getting a 5.

Applying the formula of compound probability,

Probability of getting a 2 **or** a 5,

P(2 or 5) = P(2) + P(5) – P(2 and 5)

==> 1/6 + 1/6 – 0

==> 2/6 = 1/3.

Example 2: Consider the example of finding the probability of selecting a black card or a 6 from a deck of 52 cards.

Solution:

We need to find out P(B or 6)

Probability of selecting a black card = 26/52

Probability of selecting a 6 = 4/52

Probability of selecting both a black card and a 6 = 2/52

P(B or 6) = P(B) + P(6) – P(B and 6)

= 26/52 + 4/52 – 2/52

= 28/52

= 7/13.

Conditional probability:

Conditional probability is calculating the probability of an event given that another event has already occured .

The formula for conditional probability P(A|B), read as P(A given B) is

P(A|B) = P (A and B) / P(B)

Consider the following example:

Example: In a class, 40% of the students study math and science. 60% of the students study math. What is the probability of a student studying science given he/she is already studying math?

Solution

P(M and S) = 0.40

P(M) = 0.60

P(S|M) = P(M and S)/P(S) = 0.40/0.60 = 2/3 = 0.67

Complement of an event

A complement of an event A can be stated as that which does NOT contain the occurrence of A.

A complement of an event is denoted as $P(A^c)$ or P(A’).

$$P(A^c) = 1 – P(A)$$

or it can be stated, $P(A)+P(A^c) = 1$

For example,

if A is the event of getting a head in coin toss, A^c is not getting a head i.e., getting a tail.

if A is the event of getting an even number in a die roll, A^c is the event of NOT getting an even number i.e., getting an odd number.

if A is the event of randomly choosing a number in the range of -3 to 3, A^c is the event of choosing every number that is NOT negative i.e., 0,1,2 & 3 (0 is neither positive or negative).

Consider the following example:

Example: A single coin is tossed 5 times. What is the probability of getting at least one head?

Solution:

Consider solving this using complement.

Probability of getting no head = P(all tails) = 1/32

P(at least one head) = 1 – P(all tails) = 1 – 1/32 = 31/32.

Example 1: A dice is thrown 3 times. what is the probability that at least one head is obtained?

Sol: Sample space = [HHH, HHT, HTH, THH, TTH, THT, HTT, TTT]

Total number of ways = 2 × 2 × 2 = 8. Fav. Cases = 7

P (A) = 7/8

OR

P (of getting at least one head) = 1 – P (no head) ⇒ 1 – (1/8) = 7/8

Example 2: Find the probability of getting a numbered card when a card is drawn from the pack of 52 cards.

Sol: Total Cards = 52. Numbered Cards = (2, 3, 4, 5, 6, 7, 8, 9, 10) 9 from each suit 4 × 9 = 36

P (E) = 36/52 = 9/13

Example 3: There are 5 green 7 red balls. Two balls are selected one by one without replacement. Find the probability that first is green and second is red.

Sol: P (G) × P (R) = (5/12) x (7/11) = 35/132

Example 4: What is the probability of getting a sum of 7 when two dice are thrown?

Sol: Probability math - Total number of ways = 6 × 6 = 36 ways. Favorable cases = (1, 6) (6, 1) (2, 5) (5, 2) (3, 4) (4, 3) --- 6 ways. P (A) = 6/36 = 1/6

Example 5: 1 card is drawn at random from the pack of 52 cards.

(i) Find the Probability that it is an honor card.

(ii) It is a face card.

Sol: (i) honor cards = (A, J, Q, K) 4 cards from each suits = 4 × 4 = 16

P (honor card) = 16/52 = 4/13

(ii) face cards = (J,Q,K) 3 cards from each suit = 3 × 4 = 12 Cards.

P (face Card) = 12/52 = 3/13

Example 6: Two cards are drawn from the pack of 52 cards. Find the probability that both are diamonds or both are kings.

Sol: Total no. of ways = ${}^{52}C_2$

Case I: Both are diamonds = ${}^{13}C_2$

Case II: Both are kings = ${}^{4}C_2$

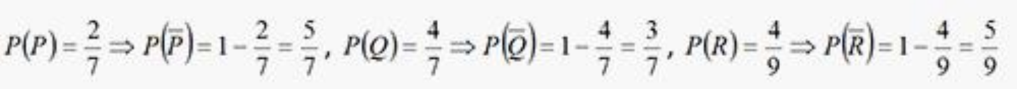
P (both are diamonds or both are kings) = $({}^{13}C_2 + {}^{4}C_2\ ) / {}^{52}C_2$

Example 7: Three dice are rolled together. What is the probability as getting at least one '4'?

Sol: Total number of ways = 6 × 6 × 6 = 216. Probability of getting number ‘4’ at least one time = 1 – (Probability of getting no number 4) = 1 – (5/6) x (5/6) x (5/6) = 91/216

Example 8: A problem is given to three persons P, Q, R whose respective chances of solving it are 2/7, 4/7, 4/9 respectively. What is the probability that the problem is solved?

Sol: Probability of the problem getting solved = 1 – (Probability of none of them solving the problem)



$$P(P)=\frac{2}{7}\Rightarrow P(\overline{P})=1-\frac{2}{7}=\frac{5}{7},\ P(Q)=\frac{4}{7}\Rightarrow P(\overline{Q})=1-\frac{4}{7}=\frac{3}{7},\ P(R)=\frac{4}{9}\Rightarrow P(\overline{R})=1-\frac{4}{9}=\frac{5}{9}$$

Probability of problem getting solved = 1 – (5/7) x (3/7) x (5/9) = (122/147)

Example 9: Find the probability of getting two heads when five coins are tossed.

Sol: Number of ways of getting two heads = ${}^{5}C_2$ = 10. Total Number of ways = 2^5 = 32

P (two heads) = 10/32 = 5/16

Example 10: What is the probability of getting a sum of 22 or more when four dice are thrown?

Sol: Total number of ways = 6^4 = 1296. Number of ways of getting a sum 22 are 6,6,6,4 = 4! / 3!

= 4
6,6,5,5 = 4! / 2!2! = 6. Number of ways of getting a sum 23 is 6,6,6,5 = 4! / 3! = 4.
Number of ways of getting a sum 24 is 6,6,6,6 = 1.
Fav. Number of cases = 4 + 6 + 4 + 1 = 15 ways. P (getting a sum of 22 or more) = 15/1296 = 5/432

Example 11: Two dice are thrown together. What is the probability that the number obtained on one of the dice is multiple of number obtained on the other dice?

Sol:Total number of cases = 6^2 = 36
Since the number on a die should be multiple of the other, the possibilities are
(1, 1) (2, 2) (3, 3) ------ (6, 6) --- 6 ways
(2, 1) (1, 2) (1, 4) (4, 1) (1, 3) (3, 1) (1, 5) (5, 1) (6, 1) (1, 6) --- 10 ways
(2, 4) (4, 2) (2, 6) (6, 2) (3, 6) (6, 3) -- 6 ways
Favorable cases are = 6 + 10 + 6 = 22. So, P (A) = 22/36 = 11/18

Example 12: From a pack of cards, three cards are drawn at random. Find the probability that each card is from different suit.
Sol: Total number of cases = $^{52}C_3$
One card each should be selected from a different suit. The three suits can be chosen in $^4C_3$ was
The cards can be selected in a total of ($^4C_3$) x ($^{13}C_1$) x ($^{13}C_1$) x ($^{13}C_1$)
Probability = $^4C_3$ x $(^{13}C_1)^3$ / $^{52}C_3$
= 4 x $(13)^3$ / $^{52}C_3$

Example 13: Find the probability that a leap year has 52 Sundays.
Sol: A leap year can have 52 Sundays or 53 Sundays. In a leap year, there are 366 days out of which there are 52 complete weeks & remaining 2 days. Now, these two days can be (Sat, Sun) (Sun, Mon) (Mon, Tue) (Tue, Wed) (Wed, Thur) (Thur, Friday) (Friday, Sat).
So there are total 7 cases out of which (Sat, Sun) (Sun, Mon) are two favorable cases. So, P (53 Sundays) = 2 / 7
Now, P(52 Sundays) + P(53 Sundays) = 1
So, P (52 Sundays) = 1 - P(53 Sundays) = 1 – (2/7) = (5/7)

Example 14: Fifteen people sit around a circular table. What are odds against two particular people sitting together?

Sol: 15 persons can be seated in 14! Ways. No. of ways in which two particular people sit together is 13! × 2!

The probability of two particular persons sitting together 13!2! / 14! = 1/7

Odds against the event = 6 : 1

Example 15: Three bags contain 3 red, 7 black; 8 red, 2 black, and 4 red & 6 black balls respectively. 1 of the bags is selected at random and a ball is drawn from it. If the ball drawn is red, find the probability that it is drawn from the third bag.

Sol: Let E1, E2, E3 and A are the events defined as follows.

E1 = First bag is chosen

E2 = Second bag is chosen

E3 = Third bag is chosen

A = Ball drawn is red

Since there are three bags and one of the bags is chosen at random, so P (E1) = P(E2) = P(E3) = 1 / 3

If E1 has already occurred, then first bag has been chosen which contains 3 red and 7 black balls. The probability of drawing 1 red ball from it is 3/10. So, P (A/E_1) = 3/10, similarly $P(A/E_2)$ = 8/10, and $P(A/E_3)$ = 4/10. We are required to find $P(E_3/A)$ i.e. given that the ball drawn is red, what is the probability that the ball is drawn from the third bag by Baye’s rule

$$= \frac{\frac{1}{3}\times\frac{4}{10}}{\frac{1}{3}\times\frac{3}{10}+\frac{1}{3}\times\frac{8}{10}+\frac{1}{3}\times\frac{4}{10}} = \frac{4}{15}.$$

Derivation of Bayes Theorem:

Statement:Let $E_1, E_2, \ldots, E_n\}$ be a set of events associated with a sample space S, where all the events $E_1, E_2, \ldots, E_n$ have nonzero probability of occurrence and they form a partition of S. Let A be any event associated with S, then according to Bayes theorem,

$$P(E_i \mid A) = \frac{P(E_i)P(E_i \mid A)}{\sum_{k=0}^{n} P(E_k)P(A|E_k)}$$

Proof:According to conditional probability formula,

$$P(E_i \mid A) = \frac{P(E_i \cap A)}{P(A)} \cdots\cdots\cdots\cdots (1)$$

Using multiplication rule of probability,

$$P(E_i \cap A) = P(E_i)P(E_i \mid A) \cdots\cdots\cdots\cdots (2)$$

Using total probability theorem,

$$P(A) = \sum_{k=0}^{n} P(E_k)P(A|E_k) \cdots\cdots\cdots\cdots (3)$$

Putting the values from equations (2) and (3) in equation 1, we get

$$P(E_i \mid A) = \frac{P(E_i)P(E_i \mid A)}{\sum_{k=0}^{n} P(E_k)P(A|E_k)}$$

Examples:

Some illustrations will improve the understanding of the concept.

Example 1:Bag I contains 4 white and 6 black balls while another Bag II contains 4 white and 3 black balls. One ball is drawn at random from one of the bags and it is found to be black. Find the probability that it was drawn from Bag I.

Solution:Let E_1 be the event of choosing the bag I, E_2 the event of choosing the bag II and A be the event of drawing a black ball.

Then,$P(E_1) = P(E_2) = \frac{1}{2}$

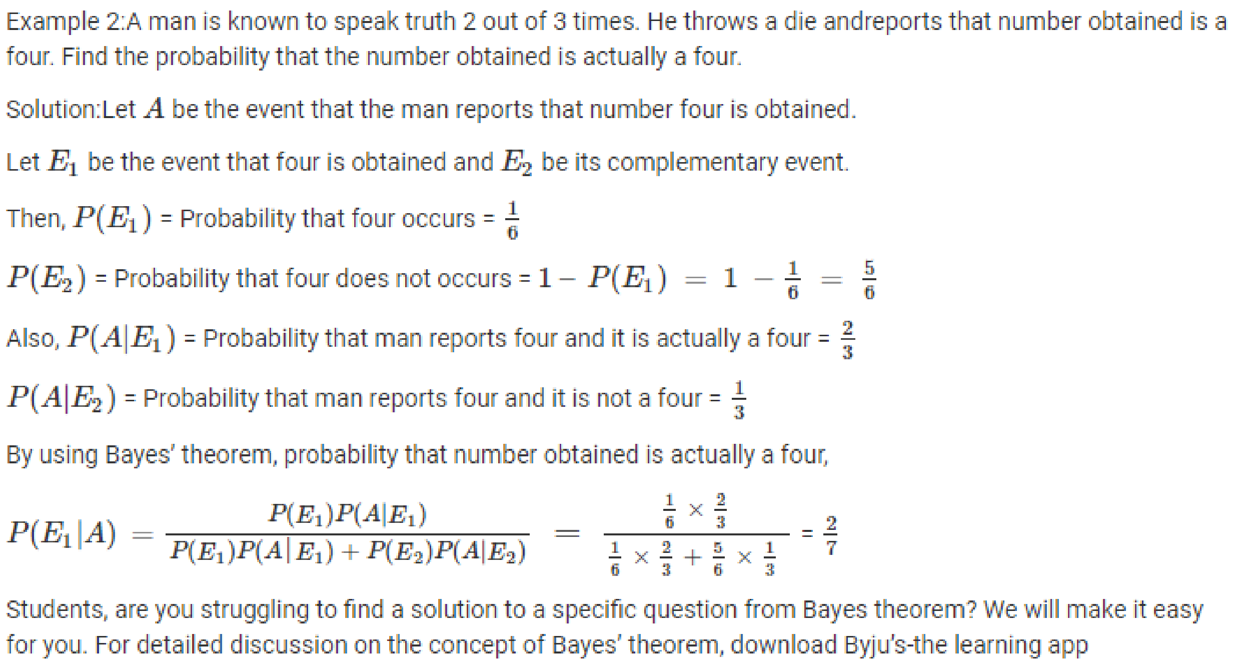
Also,$P(A|E_1) = P$(drawing a black ball from Bag I) = $\frac{6}{10} = \frac{3}{5}$

$P(A|E_2) = P$(drawing a black ball from Bag II) = $\frac{3}{7}$

By using Bayes' theorem, the probability of drawing a black ball from bag I out of two bags,

$$P(E_1|A) = \frac{P(E_1)P(A|E_1)}{P(E_1)P(A|E_1)+P(E_2)P(A|E_2)}$$

$$= \frac{\frac{1}{2} \times \frac{3}{5}}{\frac{1}{2} \times \frac{3}{7} + \frac{1}{2} \times \frac{3}{5}} = \frac{7}{12}$$



Example 2:A man is known to speak truth 2 out of 3 times. He throws a die andreports that number obtained is a four. Find the probability that the number obtained is actually a four.

Solution:Let A be the event that the man reports that number four is obtained.

Let E_1 be the event that four is obtained and E_2 be its complementary event.

Then, $P(E_1)$ = Probability that four occurs = $\frac{1}{6}$

$P(E_2)$ = Probability that four does not occurs = $1 - P(E_1) = 1 - \frac{1}{6} = \frac{5}{6}$

Also, $P(A|E_1)$ = Probability that man reports four and it is actually a four = $\frac{2}{3}$

$P(A|E_2)$ = Probability that man reports four and it is not a four = $\frac{1}{3}$

By using Bayes' theorem, probability that number obtained is actually a four,

$$P(E_1|A) = \frac{P(E_1)P(A|E_1)}{P(E_1)P(A|E_1) + P(E_2)P(A|E_2)} = \frac{\frac{1}{6} \times \frac{2}{3}}{\frac{1}{6} \times \frac{2}{3} + \frac{5}{6} \times \frac{1}{3}} = \frac{2}{7}$$

Students, are you struggling to find a solution to a specific question from Bayes theorem? We will make it easy for you. For detailed discussion on the concept of Bayes' theorem, download Byju's-the learning app

Random Variables

- A random variable X on a sample space S is a function X : S → R from S onto the set of real numbers R, which assigns a real number X (s) to each sample point 's' of S.
- Random variables (r.v.) bare denoted by the capital letters X,Y,Z,etc..
- Random variable is a single valued function.
- Sum, difference, product of two random variables is also a random variable. Finite linear combination of r.v is also a r.v .Scalar multiple of a random variable is also random variable.
- A random variable, which takes at most a countable number of values, it is called a discrete r.v. In other words, a real valued function defined on a discrete sample space is called discrete r.v.
- A random variable X is said to be continuous if it can take all possible values between certain limits .In other words, a r.v is said to be continuous when it's different values cannot be put in 1-1 correspondence with a set of positive integers.
- A continuous r.v is a r.v that can be measured to any desired degree of accuracy. Ex : age , height, weight etc..
- Discrete Probability distribution: Each event in a sample has a certain probability of occurrence. A formula representing all these probabilities which a discrete r.v. assumes is known as the discrete probability distribution.
- The probability function or probability mass function (p.m.f) of a discrete random variable X is the function f(x) satisfying the following conditions.

i) $f(x) \geq 0$

ii) $\sum_x f(x) = 1$

iii) $P(X = x) = f(x)$

- Cumulative distribution or simply distribution of a discrete r.v. X is F(x) defined by $F(x) = P(X \leq x) = \sum_{t \leq x} f(t)$ for $-\infty < x < \infty$

- If X takes on only a finite no. of values $x_1, x_2, \ldots\ldots x_n$ then the distribution function is given by

$$F(x) = \begin{cases} 0 & -\infty < x < x_1 \\ f(x_1) & x_1 \leq x < x_2 \\ f(x_1)+f(x_2) & x_2 \leq x < x_3 \\ \ldots\ldots\ldots & \\ f(x_1)+f(x_2)+\ldots..+f(x_n) & x_n \leq x < \infty \end{cases}$$

$F(-\infty) = 0$, $F(\infty)=1$, $0 \leq F(x) \leq 1$, $F(x) \leq F(y)$ if $x<y$

$P(x_k) = P(X = x_k) = F(x_k) - F(x_{k-1})$

- For a continuous r.v. X, the function f(x) satisfying the following is known as the probability density function(p.d.f.) or simply density function:

  i) $f(x) \geq 0$,$-\infty < x < \infty$

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

  iii) $P(a<X<b) = \int_a^b f(x)dx$ = Area under f(x) between ordinates x=a and x=b

$P(a<X<b) = P(a \leq x < b) = P(a < X \leq b) = P(a \leq X \leq b)$

(i.e) In case of continuous it does not matter whether we include the end points of the interval from a to b.This result in general is not true for discrete r.v.

- Probability at a point P(X=a) = $\int_{a-\Delta x}^{a+\Delta x} f(x)dx$

- Cumulative distribution for a continuous r.v. X with p.d.f. f(x), the cumulative distribution F(x) is defined as

  F(x)= P(X≤x)= $\int_{-\infty}^{\infty} f(t)dt$ -∞<x<∞

  It follows that F(-∞) = 0 , F(∞)=1, 0≤F(x)≤1 for -∞<x<∞

  f(x)= d/dx(F(x))= $F^1(x)$≥0 and P(a < x < b)= F(b)-F(a)

- In case of discrete r.v. the probability at a point i.e., P(x=c) is not zero for some fixed c however in case of continuous random variables the probability at appoint is always zero. I.e., P(x=c) = 0 for all possible values of c.

- P(E) = 0 does not imply that the event E is null or impossible event.

- If X and Y are two discrete random variables the joint probability function of X and Y is given by P(X=x,Y=y) = f(x,y) and satisfies

  (i) f(x,y) ≥ 0 (ii) $\sum_x \sum_y f(x,y)$ = 1

  The joint probability function for X and Y can be reperesented by a joint probability table.

  Table

X Y	**y_1**	**y_2**	**……**	**y_n**	**Totals**
x_1	**$f(x_1,y_1)$**	**$f(x_1,y_2)$**	**……..**	**$f(x_1,y_n)$**	**$f_1(x_{1)}$ $=P(X=x_1)$**
x_2	**$F(x_2,y_1)$**	**$f(x_2,y_2)$**	**……..**	**$f(x_2,y_n)$**	**$f_1(x_{2)}$ $=P(X=x_2)$**

……..	**…….**	**………**	**………**	**………**	**……..**
x_m	**$f(x_m,y_1)$**	**$f(x_m,y_2)$**	**…….**	**$f(x_m,y_n)$**	**$f_1(x_{m)}$ $=P(X=x_m)$**
Totals	**$f_2(y_1)$ $=P(Y=y_1)$**	**$f_2(y_2)$ $=P(Y=y_2)$**	**……..**	**$f_2(y_n)$ $=P(Y=y_n)$**	**1**

The probability ofX = x_j is obtained by adding all entries in arrow corresponding to X = x_j

Similarly the probability of Y = y_k is obtained by all entries in the column corresponding to Y =y_k

$f_1(x)$ and $f_2(y)$ are called marginal probability functions of X and Y respectively.

The joint distribution function of X and Y is defined by F(x,y)= P(X≤x,Y≤y)= $\sum_{u\le x}\sum_{v\le y} f(u,v)$

- If X and Y are two continuous r.v.'s the joint probability function for the r.v.'s X and Y is defined by

  (i) $f(x,y) \ge 0$ (ii) $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)dxdy$ =1

- P(a < X < b, c< Y < d) = $\int_{x=a}^{b}\int_{y=c}^{d} f(x,y)dxdy$

- The joint distribution function of X and Y is F(x,y) = P(X ≤ x,Y ≤ y)= $\int_{u=-\infty}^{\infty}\int_{v=-\infty}^{\infty} f(u,v)dudv$

- $\frac{\partial^2 F}{\partial x \partial y} = f(x,y)$

The Marginal distribution function of X and Y are given by P(X ≤ x) = F_1(x)= E(X)=

$$\begin{cases} \sum_i x_i f(x_i) \quad X \quad is \quad discrete \end{cases}$$

$$\begin{cases} \int_{-\infty}^{\infty} xf(x)dx \quad X \quad is\ Continuous \end{cases}$$

- $\int_{u=-\infty}^{\infty} \int_{v=-\infty}^{\infty} f(u,v)dudv$ and P(Y ≤ y) = F_2(y) = $\int_{u=-\infty}^{\infty} \int_{v=-\infty}^{\infty} f(u,v)dudv$
- The marginal density function of X and Y are given by

  f_1(x) = $\int_{v=-\infty}^{\infty} f(x,v)dv$ and f_2(y) = $\int_{u=-\infty}^{\infty} f(u,y)du$
- Two discrete random variables X and Y are independent iff

  P(X = x,Y = y) = P(X = x)P(Y = y) ∀ x,y (or)

  f(x,y) = f_1(x)f_2(y) ∀ x, y
- Two continuous random variables X and Y are independent iff

  P(X ≤ x,Y ≤ y) = P(X ≤ x)P(Y ≤ y) ∀ x,y (or)

  f(x,y) = f_1(x)f_2(y) ∀ x, y

If X and Y are two discrete r.v. with joint probability function f(x,y) then

P(Y = y|X=x) = $\frac{f(x,y)}{f_1(x)}$ = f(y|x)

Similarly, P(X = x|Y=y) = $\frac{f(x,y)}{f_2(y)}$ = f(x|y)

If X and Y are continuous r.v. with joint density function f(x,y) then $\frac{f(x,y)}{f_1(x)}$ = f(y|x) and

$\frac{f(x,y)}{f_2(y)}$ = f(x|y)

Expectation or mean or Expected value : The mathematical expectation or expected value of r.v. X is denoted by E(x) or μ and is defined as

- If X is a r.v. then E[g(X)] = $\begin{cases} \sum_{x} g(x)f(x) & \text{for Discrete} \\ \int_{-\infty}^{\infty} g(x)f(x)dx & \text{For Continuous} \end{cases}$

- If X, Y are r.v.'s with joint probability function f(x,y) then

E[g(X,Y)] = $\begin{cases} \sum_{x}\sum_{y} g(x,y)f(x,y) & \text{for discrete r.v.'s} \\ \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y)f(x,y)dxdy & \text{for continuous r.v.'s} \end{cases}$

If X and Y are two continuous r.v.'s the joint density function f(x,y) the conditional expectation or the conditional mean of Y given X is E(Y |X = x) = $\int_{-\infty}^{\infty} yf(y \mid x)dy$

Similarly, conditional mean of X given Y is E(X |Y = y) = $\int_{-\infty}^{\infty} xf(x \mid y)dx$

- Median is the point, which divides the entire distribution into two equal parts. In case of continuous distribution median is the point, which divides the total area into two equal parts. Thus, if M is the median then $\int_{-\infty}^{M} f(x)dx = \int_{M}^{\infty} f(x)dx$ =1/2. Thus, solving any one of the equations for M we get the value of median. Median is unique

- Mode: Mode is the value for f(x) or P(x_i) at attains its maximum

  For continuous r.v. X mode is the solution of $f^1(x) = 0$ and $f^{11}(x) < 0$

  provided it lies in the given interval. Mode may or may not be unique.

- Variance: Variance characterizes the variability in the distributions with same mean can still have different dispersion of data about their means

Variance of r.v. X denoted by Var(X) and is defined as

$$\text{Var}(X) = E\left[\,(X - \mu\ )^2\ \right] = \begin{cases} \sum_{x}(x-\mu)^2 f(x) & \text{for discrete} \\ \int_{-\infty}^{\infty}(x-\mu)^2 f(x)dx & \text{for continuous} \end{cases}$$

where $\mu = E(X)$

- If c is any constant then E(cX) = c E(X)
- If X and Y are two r.v.'s then E(X+Y) = E(X)+E(Y)
- IF X,Y are two independent r.v.'s then E(XY) = E(X)E(Y)
- If $X_1,X_2,\text{-------},X_n$ are random variables then $E(c_1X_1 + c_2X_2 + \text{------} + c_nX_n) = c_1E(X_1)+c_2E(X_2)+\text{-----}+c_nE(X_n)$ for any scalars $c_1,c_2,\text{------},c_n$ If all expectations exists
- If $X_1,X_2,\text{-------},X_n$ are independent r.v's then $E\left(\prod_{i=1}^{n} X_i\right) = \prod_{i=1}^{n} E(X_i)$ if all expectations exists.
- Var (X) = E (X^2) –[E (X)$]^2$
- If 'c' is any constant then var (cX) = c^2var(X)
- The quantity $E[(X-a)^2]$ is minimum when a =μ= E(X)
- If X and Y are independent r.v.'s then Var(X ± Y) = Var(X) ± Var(Y)

Module-II

PROBABILITY DISTRIBUTION

Binomial Distribution

- A random variable X is said to follow binomial distribution if it assumes only non-negative

  values and its probability mass function is given by

  $$P(X = x) = P(x) = \binom{n}{x} p^x q^{n-x} \quad \text{where} \quad x = 0,1,2,3,\ldots.n \quad q = 1\text{-}p$$

  where n, p are known as parameters, n- number of independent trials p- probability of success in each

  trial, q- probability of failure.

- Binomial distribution is a discrete distribution.
- The notation X ~B(n,p) is the random variable X which follows the binomial distribution with parameters n and p
- If n trials constitute an experiment and the experiment is repeated N times the frequency function of the binomial distribution is given by f(x) = NP(x). The expected frequencies of 0,1,2,….. n successes are the successive terms of the binomial expansion $N(p+q)^n$
- The mean and variance of Binomial distribution are np , npq respectively.
- **Mode of the Binomial distribution**: Mode of B.D. Depending upon the values of (n+1)p

  (i) If (n+1)p is not an integer then there exists a unique modal value for binomial distribution and it is 'm'= integral part of (n+1)p

  (ii) If (n+1)p is an integer say m then the distribution is Bi-Modal and the two modal values are m and m-1

- Moment generating function of Binomial distribution: If X ~B(n,p)then $M_X(t)=(q+pe^t)^n$
- The sum of two independent binomial variates is not a binomial varaite. In other words, Binomial distribution does not posses the additive or reproductive property.
- For B.D. $\gamma_1 = \sqrt{\beta_1} = \dfrac{1-2p}{\sqrt{npq}}$ $\gamma_2 = \beta_2 - 3 = \dfrac{1-6pq}{npq}$
- If $X_1 \sim B(n_1,p)$ and $X_2 \sim B(n_2,p)$ then $X_1+X_2 \sim B(n_1+n_2,p)$.Thus the B.D. Possesses the additive or reproductive property if $p_1=p_2$

Poisson Distribution

- Poisson Distribution is a limiting case of the Binomial distribution under the following conditions:

(i) n, the number of trials is infinitely large.

(ii) P, the constant probability of success for each trial is indefinitely small.

(iii) np= λ, is finite where λ is a positive real number.

- A random variable X is said to follow a Poisson distribution if it assumes only non-negative values and

its p.m.f. is given by

$$P(x,\lambda)= P(X= x) =: \begin{cases} \frac{e^{-\lambda}\lambda^{x}}{x!} & x= 0,1,2,3,\ldots\ \lambda > 0 \\ 0 & \text{Other wise} \end{cases}$$

Here λ is known as the parameter of the distribution.

- We shall use the notation X~ P(λ) to denote that X is a Poisson variate with parameter λ
- Mean and variance of Poisson distribution are equal to λ.
- The coefficient of skewness and kurtosis of the poisson distribution are $\gamma_1 = \sqrt{\beta_1}= 1/\sqrt{\lambda}$ and $\gamma_2= \beta_2-3=1/\lambda$. Hence the poisson distribution is always a skewed distribution. Proceeding to limit as λ tends to infinity we get $\beta_1 = 0$ and $\beta_2=3$
- Mode of Poisson Distribution: Mode of P.D. Depending upon the value of λ

(i) when λ is not an integer the distribution is uni- modal and integral part of λ is the unique modal value.

(ii) When λ = k is an integer the distribution is bi-modal and the two modals are k-1 and k.

- Sum of independent poisson variates is also poisson variate.
- The difference of two independent poisson variates is not a poisson variate.
- **Moment generating function of the P.D.**

If X~ P(λ) then $M_X(t) = e^{\lambda(e^t-1)}$

- Recurrence formula for the probabilities of P.D. (Fitting of P.D.)

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

- Recurrence relation for the probabilities of B.D. (Fitting of B.D.)

  P(x+1) = $\left\{\frac{n-x}{x+1}.\frac{p}{q}\right\}p(x)$

Normal Distribution

- A random variable X is said to have a normal distribution with parameters μ called mean and σ^2 called variance if its density function is given by the probability law

  f(x; μ, σ) = $\frac{1}{\sigma\sqrt{2\pi}}\exp\left[\frac{-1}{2}\left\{\frac{x-\mu}{\sigma}\right\}^2\right]$, -∞< x <∞, -∞<μ<∞, σ> 0

- A r.v. X with mean μ and variance σ^2 follows the normal distribution is denoted by

  X~ N(μ, σ^2)

- If X~ N(μ, σ^2) then Z = $\frac{X-\mu}{\sigma}$ is a standard normal variate with E(Z) = 0 and var(Z)=0 and we write Z~ N(0,1)

- The p.d.f. of standard normal variate Z is given by f(Z) = $\frac{1}{\sqrt{2\pi}}e^{-z^2/2}$, -∞< Z<∞

- The distribution function F(Z) = P(Z ≤ z) = $\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z}e^{-t^2/2}dt$

- F(-z) = 1 – F(z)

- P(a < z ≤ b) = P(a ≤ z < b)= P(a <z < b)= P(a ≤ z ≤ b)= F(b) – F(a)

- If X~ N(μ, σ^2) then Z = $\frac{X-\mu}{\sigma}$ then P(a ≤ X ≤ b) = $F\left(\frac{b-\mu}{\sigma}\right) - F\left(\frac{a-\mu}{\sigma}\right)$

- N.D. is another limiting form of the B.D. under the following conditions:

  i) n , the number of trials is infinitely large.

ii) Neither p nor q is very small

- **Chief Characteristics of the normal distribution and normal probability curve:**

i) The curve is bell shaped and symmetrical about the line $x = \mu$

ii) Mean median and mode of the distribution coincide.

iii) As x increases numerically f(x) decreases rapidly.

iv) The maximum probability occurring at the point $x= \mu$ and is given by

$[P(x)]_{max} = 1/\sigma\sqrt{2\Pi}$

v) $\beta_1 = 0$ and $\beta_2 = 3$

vi) $\mu_{2r+1} = 0$ (r = 0,1,2……) and $\mu_{2r} = 1.3.5….(2r-1)\sigma^{2r}$

vii) Since f(x) being the probability can never be negative no portion of the curve lies below x- axis.

viii) Linear combination of independent normal variate is also a normal variate.

ix) X- axis is an asymptote to the curve.

x) The points of inflexion of the curve are given by $x = \mu\pm\sigma$, $f(x) = \dfrac{1}{\sigma\sqrt{2\pi}} e^{-1/2}$

xi) Q.D. **:** M.D.**:** S.D. **::** $\frac{2}{3}\sigma$ **:** $\frac{4}{5}\sigma$ **:** σ**::** $\frac{2}{3}$ **:** $\frac{4}{5}$ **:** 1 Or Q.D. **:** M.D.**:** S.D. **::10:12:15**

xii) Area property: $P(\mu- \sigma< X <\mu + \sigma) = 0.6826 = P(-1 < Z < 1)$

$P(\mu- 2\sigma< X <\mu + 2\sigma) = 0.9544 = P(-2 < Z < 2)$

$P(\mu- 3\sigma< X <\mu +3 \sigma) = 0.9973 = P(-3 < Z < 3)$

$P(|Z| > 3) = 0.0027$

- m.g.f. of N.D. If $X\sim N(\mu, \sigma^2)$ then $M_X(t) = e^{\mu t} +t^2\sigma^2/2$

IfZ~ N(0,1) then $M_Z(t) = e^{t^2/2}$

Continuity Correction:

- The N.D. applies to continuous random variables. It is often used to approximate distributions of discrete r.v. Provided that we make the continuity correction.

- If we want to approximate its distribution with a N.D. we must spread its values over a continuous scale. We do this by representing each integer k by the interval from k-1/2 to k+1/2 and at least k is represented by the interval to the right of k-1/2 to at most k is represented by the interval to the left of k+1/2.

- **Normal approximation to the B.D:**

  X~ B(n, p) and if Z = $\frac{X - np}{\sqrt{np(1-p)}}$ then Z ~ N(0,1) as n tends to infinity and F(Z) =

  F(Z)= P(Z ≤ z) = $\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-t^2/2} dt$ -∞< Z <∞

- Use the normal approximation to the B.D. only when (i) np and n(1-p) are both greater than 15 (ii) n is small and p is close to ½

- **Poisson process:** Poisson process is a random process in which the number of events (successes) x occurring in atime interval of length T is counted. It is continuous parameter, discrete stable process. By dividing T into n equal parts of length Δt we have T = n **.**ΔT. Assuming that (i) P ∝ΔT or P = αΔt (ii) The occurrence of events are independent (iii) The probability of more than one substance during a small time interval Δt is negligible.

  As n →∞**,** the probability of x success during a time interval T follows the P.D. with parameter λ = np = α**T** where α is the average(mean) number of successes for unit time.

  <u>PROBLEMS:</u>

1:A random variable x has the following probability function:

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

Find (i) k (ii) P(x<6) (iii) P(x>6)

Solution:

(i) since the total probability is unity, we have $\sum_{x=0}^{n} p(x) = 1$

i.e., 0 + k +2k+ 2k+ 3k+ k^2+7k^2+k=1

i.e., 8k^2+9k-*1=0*

k=1,-1/8

(ii) P(x<6)= 0 + k +2k+ 2k + 3k

=1+2+2+3=8

iii) P(x>6)= k^2 +7k^2+k

=9

2. Let X denotes the minimum of the two numbers that appear when a pair of fair dice is thrown once. Determine (i) Discrete probability distribution (ii) Expectation (iii) Variance

Solution:

When two dice are thrown, total number of outcomes is 6x6-36

In this case, sample space S=

$$\begin{matrix} \{(1,1)(1,2)(1,3)(1,4)(1,5)(1,6) \\ (2,1)(2,2)(2,3)(2,4)(2,5)(2,6) \\ (3,1)(3,2)(3,3)(3,4)(3,5)(3,6) \\ (4,1)(4,2)(4,3)(4,4)(4,5)(4,6) \\ (5,1)(5,2)(5,3)(5,4)(5,5)(5,6) \\ (6,1)(6,2)(6,3)(6,4)(6,5)(6,6) \end{matrix}$$

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

$$\left\{\begin{matrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 2 & 2 & 2 & 2 \\ 1 & 2 & 3 & 3 & 3 & 3 \\ 1 & 2 & 3 & 4 & 4 & 4 \\ 1 & 2 & 3 & 4 & 5 & 5 \\ 1 & 2 & 3 & 4 & 5 & 6 \end{matrix}\right\}$$

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

For minimum 1, the favorable cases are 11

Therefore, P(x=1)=11/36

P(x=2)=9/36, P(x=3)=7/36, P(x=4)=5/36, P(x=5)=3/36, P(x=6)=1/36

The probability distribution is

X	1	2	3	4	5	6
P(x)	11/36	9/36	7/36	5/36	3/36	1/36

(ii)Expectation mean = $\sum p_i x_i$

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

Or $\mu = \frac{1}{36}[11 + 8 + 21 + 20 + 15 + 6] = \frac{9}{36} = 2.5278$

(ii) variance = $\sum p_i x^2{}_i - \mu^2$

$$E(x) = \frac{11}{36}1 + \frac{9}{36}4 + \frac{7}{36}9 + \frac{5}{36}16 + \frac{3}{36}25 + \frac{1}{36}36 - (2.5278)^2$$

=1.9713

3: A continuous random variable has the probability density function

$$f(x) = \begin{cases} kxe^{-\lambda x},\ for\ x \geq 0, \lambda > 0 \\ 0, otherwise \end{cases}$$

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

Solution:

(i) since the total probability is unity, we have $\int_{-\infty}^{\infty} f(x)dx = 1$

$$\int_{-\infty}^{0} 0dx + \int_{0}^{\infty} kxe^{-\lambda x} dx = 1$$

i.e., $\int_{0}^{\infty} kxe^{-\lambda x} dx = 1$

$$k\left[x\left(\frac{e^{-\lambda x}}{-\lambda}\right) - 1\left(\frac{e^{-\lambda x}}{\lambda^2}\right)\right]_0^\infty \ or\ k = \lambda^2$$

(ii) mean of the distribution $\mu = \int_{-\infty}^{\infty} xf(x)dx$

$$\int_{-\infty}^{0} 0dx + \int_{0}^{\infty} kx^2 e^{-\lambda x} dx$$

$$\lambda^2 \left[x^2 \left(\frac{e^{-\lambda x}}{-\lambda} \right) - 2x \left(\frac{e^{-\lambda x}}{\lambda^2} \right) + 2 \left(\frac{e^{-\lambda x}}{\lambda^3} \right) \right]_0^{\infty}$$

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

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

$$\sigma^2 = \int_{-\infty}^{\infty} x^2 f(x) dx - \frac{4}{\lambda^2}$$

$$\lambda^2 \left[x^3 \left(\frac{e^{-\lambda x}}{-\lambda} \right) - 3x^2 \left(\frac{e^{-\lambda x}}{\lambda^2} \right) + 6x \left(\frac{e^{-\lambda x}}{\lambda^3} \right) - 6 \left(\frac{e^{-\lambda x}}{\lambda^4} \right) \right]_0^{\infty} - \frac{4}{\lambda^2}$$

$$= \frac{2}{\lambda^2}$$

4:

Out of 800 families with 5 children each, how many would you expect to have (i)3 boys (ii)5girls (iii)either 2 or 3 boys ? Assume equal probabilities for boys and girls

Solution

P(3boys)=P(r=3)=P(3)= $\frac{1}{2^5} {}^5C_3 = \frac{5}{16}$ per family

Thus for 800 families the probability of number of families having 3 boys= $\frac{5}{16}(800) = 250$ families

P(5 girls)=P(no boys)=P(r=0)= $\frac{1}{2^5} {}^5C_0 = \frac{1}{32}$ per family

Thus for 800 families the probability of number of families having 5girls= $\frac{1}{32}(800) = 25$ families

P(either 2 or 3 boys =P(r=2)+P(r=3)=P(2)+P(3)

$\frac{1}{2^5} {}^5C_2 + \frac{1}{2^5} {}^5C_3$=5/8 per family

Expected number of families with 2 or 3 boys = $\frac{5}{8}(800) = 500$ families.

5: Average number of accidents on any day on a national highway is 1.8. Determine the probability that the number of accidents is (i) at least one (ii) at most one

Solution:

Mean= $\lambda = 1.8$

We have P(X=x)= p(x)$\frac{e^{-\lambda}\lambda^{x}}{x!}$ =$\frac{e^{-1.8}1.8^{x}}{x!}$

P (at least one) =P(x≥1)=1-P(x=0)

=1-0.1653

=0.8347

P (at most one) =P (x≤1)

=P(x=0)+P(x=1)

= 0.4628

6: The mean weight of 800 male students at a certain college is 140kg and the standard deviation is 10kg assuming that the weights are normally distributed find how many students weigh I) Between 130 and 148kg ii) more than 152kg

Solution:

Let μ be the mean and σ be the standard deviation. Then μ =140kg and σ =10pounds

(i) When x= 138, $z = \frac{x-\mu}{\sigma} = \frac{138-140}{10} = -0.2 = z_1$

When x= 138, $z = \frac{x-\mu}{\sigma} = \frac{148-140}{10} = 0.8 = z_2$

$\therefore$ P(138≤x≤148)=P(-0.2≤z≤0.8)

=A(z_2)+A(z_1)

=A(0.8)+A(0.2)=0.2881+0.0793=0.3674

Hence the number of students whose weights are between 138kg and 140kg =0.3674x800=294

(ii) When x=152,$\frac{x-\mu}{\sigma}$=$\frac{152-140}{10} = 1.2$=z_1

Therefore P(x>152)=P(z>z$_1$)=0.5-A(z$_1$)

=0.5-0.3849=0.1151

Therefore number of students whose weights are more than 152kg =800x0.1151=92.

Exercise Problems:

1. Two coins are tossed simultaneously. Let X denotes the number of heads then find i) E(X) ii) E(X^2) iii)E(X^3) iv) V(X)
2. If f(x)=k$e^{-|x|}$ is probability density function in the interval, $-\infty < x < \infty$, then find i) k ii) Mean iii) Variance iv) P(0<x<4)
3. Out of 20 tape recorders 5 are defective. Find the standard deviation of defective in the sample of 10 randomly chosen tape recorders. Find (i) P(X=0) (ii) P(X=1) (iii) P(X=2) (iv) P (1<X<4).
4. In 1000 sets of trials per an event of small probability the frequencies f of the number of x of successes are

f	0	1	2	3	4	5	6	7	Total
x	305	365	210	80	28	9	2	1	1000

Fit the expected frequencies.

5. If X is a normal variate with mean 30 and standard deviation 5. Find the probabilities that

i) P(26 ≤ X≤40) ii) P(X ≥ 45)

6. The marks obtained in Statistics in a certain examination found to be normally distributed. If

15% of the students greater than or equal to 60 marks, 40% less than 30 marks. Find the

mean and standard deviation.

7.If a Poisson distribution is such that $P(X = 1) = \frac{3}{2} P(X = 3)$ then find (i) $P(X \geq 1)$ (ii) $P(X \leq 3)$ (iii) $P(2 \leq X \leq 5)$.

. A random variable X has the following probability function:

X	-2	-1	0	1	2	3
P(x)	0.1	K	0.2	2K	0.3	K

Then find (i) k (ii) mean (iii) variance (iv) P(0 < x < 3)

Module-III
CORRELATION AND REGRESSION

- **Correlation**: In a bivariate distribution, if the change in one variable effects the change in other variable, then the variables are called correlated.
- **Covariance** between two random variables X and Y is denoted by Cov(X,Y) is defined as E(XY)-E(X)E(Y)
- If X and Y are independent then Cov(X,Y) = 0
- **Karl Pearson Correlation Coefficient** between two r.v. X and Y usually denoted by r(X,Y) or simply r_{XY} is a numerical measure of a linear relationship between them and is defined as $r = r(X,Y) = cov(X,Y)/\sigma_x\sigma_y$

  It is also called **product moment correlation coefficient.**

- If (x_i,y_i); I = 1,2…n is bivariate distribution then, then

  Cov (X,Y) = E[{X-E(X)}{Y-E(Y)}]

  $$= (1/n)\sum(x_i-\bar{x})(y_i-\bar{y}) = (1/n)\sum x_iy_i - \bar{x}\ \bar{y}$$

  $$\sigma_X^2 = E[X-E(X)]^2 = (1/n)\sum(x_i-\bar{x})^2 = (1/n)\sum x_i^2-(\bar{x})^2$$

$$\sigma_Y^2 = E[Y-E(Y)]^2 = (1/n)\sum(y_i-\bar{y})^2 = (1/n)\sum y_i^2-(\bar{y})^2$$

- Computational formula for r(X,Y) = $\dfrac{\frac{1}{n}\sum xy-\bar{x}\bar{y}}{\sqrt{\left[\left(\frac{1}{n}\sum x^2\right)-\bar{x}^2\right]}\sqrt{\left[\left(\frac{1}{n}\sum y^2\right)-\bar{y}^2\right]}}$
- $-1 \le r \le 1$
- If r = 0 then X,Y are uncorrelated.
- If r = -1 then correlation is perfect and negative.
- If r = 1then the correlation is perfect and positive.

- r is independent of change of origin and scale
- Two independent variables are uncorrelated. Converse need not be true.
- **The correlation coefficient for Bivariate frequency distribution**:

  The bivariate data on X on Y are presented in a two-way correlation table with n classes of Y placed along the horizontal lines and m classes of X along vertical lines and f_{ij} is the frequency of the individuals lying in i, j th cell.

  $\sum_x f(x,y)$ =g(y),is the sum of the frequencies along any row and

  $\sum_y f(x,y)$ =f(x),is the sum of the frequencies along any column.

  $$\sum_x \sum_y f(x,y) = \sum_y \sum_x f(x,y) = \sum_x f(x) = \sum_y g(y) = \text{N}$$

  $$\bar{x} = \frac{1}{N}\sum_x xf(x), \qquad \bar{y} = \frac{1}{N}\sum_y yg(y)$$

  $$\sigma_X^2 = \frac{1}{N}\sum_x x^2 f(x) - \bar{x}^2 \text{ and } \sigma_Y^2 = \frac{1}{N}\sum_y y^2 g(y) - \bar{y}^2$$

  $$\text{Cov(X,Y)} = \frac{1}{N}\sum_x \sum_y \; xyf(x,y) - \bar{x}\,\bar{y}$$

  $$\text{r}= \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

- **Rank Correlation**: Let (x_i,y_i) for I = 1,2,…n be the ranks of the ith individuals in the characteristics A and B respectively, Pearsonian coefficient of correlation between x_i and y_i are called rank correlation coefficient between A and B for that group of individual.
- **The Spearman's rank correlation** between the two variables X and Y takes the values 1,2…n denoted by ρ and is defined as $\rho = 1 - \frac{\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$

where $d_i = x_i - y_i$ (In general $x_i \neq y_i$)

In 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 slightly different from each other and the next item will get the rank next to the rank already assumed. The adjustment or correction is made in the rank correlation formula. In the formula we add factor $\frac{m(m^2-1)}{12}$ to $\sum d^2$, wherem is the number of times an item is repeated. This correction factor is to be added to each repeated value in both X-series and Y- series.

- $-1 \leq \rho \leq 1$
- **Regression analysis** is a mathematical measure of the average relationship between two or more variables in terms of the original units of the data.
- The variable whose value is influenced or is to be predicted is called **dependent variable** and the variable, which influences the values or is used for the prediction is called **independent variable**. Independent variable is also known as **regressor or predictor or explanatory variable** while the dependent variable is also known as **regressed or explained variable**.
- If the variables in bivariate distributions are related we will find that the points in the scatter diagram will cluster round some curve called the " curve of regression". If the curve is a straight line, it is called line of regression and there is said to be **linear Regression** between thevariables, otherwise the regression is said to be curvilinear. The line of regression is line of best fit and is obtained by principle of least squares.
- In the bivariate distribution (x_i, y_i) ; i = 1,2,….n Y is dependent variable and X is independent variable. The line of regression Y on X is Y = a + b X.

  i.e. $Y - \bar{y} = r\frac{\sigma_Y}{\sigma_X}(X - \bar{x})$

  Similarly the line of regression X on Y is X = a + b Y

  i.e.,$X - \bar{x} = r\frac{\sigma_X}{\sigma_Y}(Y - \bar{y})$

If X and Y are any random variables the two regression lines are

$$Y - E(Y) = \frac{Cov(X,Y)}{\sigma_X^2}[X - E(X)]$$

$$X - E(X) = \frac{Cov(X,Y)}{\sigma_Y^2}[Y - E(Y)]$$

- Both lines of regression passes through the point $(\bar{x}, \bar{y})$ i.e., the mean values $(\bar{x}, \bar{y})$ can be obtained at the point of intersection of regression lines.
- The slope of regression line Y on X is also called the regression coefficient Y on X. It represents the increment in the value of dependent variable Y corresponding to a unit change in the value of independent variable X. We write, b_{YX} = Regression coefficient Y on X $= r\frac{\sigma_Y}{\sigma_X}$
- Similarly the coefficient of regression of X on Y indicates the change in value of variable X corresponding to a unit change in value of variable Y and is given by b_{XY} = Regression coefficient X on Y $= r\frac{\sigma_X}{\sigma_Y}$
- Correlation Coefficient is the geometric mean between the regression coefficients.
- The sign of correlation coefficients is same as that of sign of regression coefficients
- If one of the regression coefficients is greater than unity the other must be less than unity.
- The modulus value of the arithmetic mean of regression coefficient is not less than modulus value of correlation coefficient r.
- Regression coefficients are independent of the change of origin but not scale.
- If θ is the acute angle between two lines of regression then

$$\theta = \text{Tan}^{-1}\left\{\frac{1-r^2}{|r|}\left(\frac{\sigma_X\sigma_Y}{\sigma_X^2+\sigma_Y^2}\right)\right\}$$

If r = 0 then variables X and Y are uncorrelated. The lines of regressions are Y = $\bar{y}$ and X = $\bar{x}$ which are perpendicular to each other and are parallel to x- axis and y-axis respectively.

If r = ± 1 , the two lines of regression coincide.

- **Regression Curves:** The conditional mean E (Y|X = x) for a continuous distribution is called the regression function Y on X and the graph of this function of x is known as regression curve of Y on X.
  The regression function of X on Y is E (X|Y = y) and the graph of this function of y is called regression curve (of the mean) of X on Y.

- **Multiple Regression** analysis is an extension of (simple) regression analysis in which two or more independent variables are used to estimate the value of dependent variable.

  Least square regression planes fitting of N data points (X_1,X_2,X_3) in a three dimensional scatter diagram. The least square regression plane of X_1 on X_2 and X_3 is $X_1 = a + b\,X_2 + cX_3$ where a,b,c are determined by solving simultaneously the normal equations:

$$\sum X_1 = an + b\sum X_2 + c\sum X_3$$

$$\sum X_1X_2 = a\sum X_2 + b\sum X_2^2 + c\sum X_2X_3$$

$$\sum X_1X_3 = a\sum X_3 + b\sum X_2X_3 + c\sum X_3^2$$

  Similarly for the regression plane of X_2 on X_1 and X_3 and the regression plane of X_3 on X_1 and X_2

- The linear regression equation of X_1 on X_2 ,X_3 and X_4 can be written as

  $X_1 = a + b\,X_2 + c\,X_3 + d\,X_4$

<u>PROBLEMS:</u>

1. **Calculate the coefficient of correlation from the following data**

x	**12**	**9**	**8**	**10**	**11**	**13**	**7**
y	**14**	**8**	**6**	**9**	**11**	**12**	**13**

Solution: Here $X = x - \bar{x}$ and $Y = y - \bar{y}$

$$\bar{x} = \frac{\sum x_i}{n} = \frac{12+9+8+10+11+13+7}{7} = 10$$

$$\bar{y} = \frac{\sum y_i}{n} = \frac{14+8+6+9+11+12+13}{7} = 10.4$$

x	y	$X = x - \bar{x}$	$Y = y - \bar{y}$	XY	X^2	Y^2
12 9 8 10 11 13 7	14 8 6 9 11 12 13	2 -1 -2 0 1 3 -3	3.6 -2.4 -4.4 -1.4 0.6 1.6 2.6	7.2 2.4 8.8 0 0.6 4.8 -7.8	4 1 4 0 1 9 9	12.9 5.7 19.3 1.9 0.3 2.5 6.7
				$\sum XY = 16$	$\sum X^2 = 28$	$\sum Y^2 = 49.3$

$\therefore$ Correlation Coefficient $r = \dfrac{\sum XY}{\sqrt{\sum X^2 . \sum Y^2}}$

$$= \frac{16}{\sqrt{28 \times 49.3}}$$

$\therefore r = 0.43$

$\therefore r$ is positive.

2. **The ranks of 16 students in Mathematics and Statistics are as follows (1,1),(2,10),(3,3),(4,4),(5,5),(6,7),(7,2),(8,6),(9,8),(10,11),(11,15),(12,9),(13,14),(1**

4,12),(15,16),(16,13). Calculate the rank correlation coefficient for proficiencies of this group in mathematics and statistics.

Solution:

Ranks in Mathematics (X)	**Ranks in Statistics** (Y)	$D = X - Y$	D^2
1	1	0	0
2	10	-8	64
3	3	0	0
4	4	0	0
5	5	0	0
6	7	-1	1
7	2	5	25
8	6	2	4
9	8	1	1
10	11	-1	1
11	15	-4	16
12	9	3	9
13	14	-1	1
14	12	2	4
15	16	-1	1
16	13	3	9
			$\sum D^2 = 136$

$\therefore$ **Rank Correlation Coefficient** $\rho = 1 - \dfrac{6\sum D^2}{N(N^2 - 1)}$

$$= 1 - \frac{6 \times 136}{16 \times 225}$$

$$\therefore \rho = 0.8$$

Problem:

Determine the regression equation which best fit to the following data:

x	10	12	13	16	17	20	25
y	10	22	24	27	29	33	37

Solution: The regression equation of y on x is $y = a + bx$

The normalequations are

$$\sum y = na + b\sum x$$

$$\sum xy = a\sum x + b\sum x^2$$

x	y	x^2	xy
10	10	100	100
12	22	144	264
13	24	169	312
16	27	256	432
17	29	289	493
20	33	400	660
25	37	625	925
$\sum x = 113$	$\sum y = 182$	$\sum x^2 = 1983$	$\sum xy = 3186$

Substitute the above values in normal equations

$$\sum y = na + b\sum x \Rightarrow 182 = 7a + 113b - - - - - (1)$$

$$\sum xy = a\sum x + b\sum x^2 \Rightarrow 3186 = 113a + 1983b - - - - - (2)$$

Solve equations (1) and (2) we get

$a = 0.7985$ and $b = 1.5611$

Now substitute a, b values in regression equation

$\therefore$ The regression equation of y on x is $y = 0.7985 + 1.5611x$

3. Give the following data compute multiple coefficient of correlation of X_3 on X_1 and X_2.

X_1	3	5	6	8	12	14
X_2	16	10	7	4	3	2
X_3	90	72	54	42	30	12

Solution: here $n = 6, \quad \overline{X}_1 = \frac{48}{6} = 8, \quad \overline{X}_2 = \frac{42}{6} = 7, \quad \overline{X}_3 = \frac{300}{6} = 50$

Now we calculate values of r_{12}, r_{13} and r_{23}

	$x_1 = X_1 - \overline{X}_1$			$x_2 = X_2 - \overline{X}_2$			$x_3 = X_3 - \overline{X}_3$					
S.NO	X_1	x_1	$x_1^{\;2}$	X_2	x_2	x_2^2	X_3	x_3	$x_3^{\;2}$	x_1x_2	x_2x_3	x_3x_1
1	3	-5	25	16	9	81	90	40	1600	-45	360	-200
2	5	-3	9	10	3	9	72	22	484	-9	66	-66
3	6	-2	4	7	0	0	54	4	16	0	0	-8
4	8	0	0	4	-3	9	42	-8	64	0	24	0
5	12	4	16	3	-4	16	30	-20	400	-16	80	-80
6	14	6	36	2	-5	25	12	-38	1444	-30	190	-228
	68	0	90	42	0	140	300	0	4008	-100	-582	720

$$r_{12} = \frac{\sum x_1x_2}{\sqrt{\sum x_1^2 \sum x_2^2}} = \frac{-100}{\sqrt{90 \times 140}} = -0.89$$

$$r_{12} = \frac{\sum x_1x_3}{\sqrt{\sum x_1^2 \sum x_3^2}} = \frac{-582}{\sqrt{90 \times 4008}} = -0.97$$

$$r_{12} = \frac{\sum x_2x_3}{\sqrt{\sum x_2^2 \sum x_3^2}} = \frac{720}{\sqrt{140 \times 4008}} = 0.96$$

$$R_{3.12} = \sqrt{\frac{r_{13}^2 + r_{23}^2 - 2r_{13}r_{23}r_{12}}{1 - r_{12}^2}} = 0.987$$

Exercise Problems:

1. Calculate the Karl Pearson’s coefficient of correlation from the following data

x	15	18	20	24	30	35	40	50
y	85	93	95	105	120	130	150	160

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

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

3. Find the most likely production corresponding to a rainfall 40 from the following data:

	Rain fall(X)	**Production(Y)**
Average	30	500Kgs
Standard deviation	5	100Kgs
Coefficient of correlation	0.8	

4. If θ is the angle between two regression lines and S.D. of Y is twice the S.D. of X and r=0.25,

5. find tan θ The joint probability density function $f(x,y)=\begin{cases} Ae^{-x-y}, 0<x<y, 0<y<\infty \\ 0.\ \text{Otherwise} \end{cases}$.

   Determine A.

6. Determine the regression equation which best fit to the following data:

x	10	12	13	16	17	20	25
y	10	22	24	27	29	33	37

MODULE –IV

TEST OF HYPOTHESIS - I

Sampling Distribution

- **Population** is the set or collection or totality of the objects, animate or inanimate, actual or hypothetical under study. Thus, mainly population consists of set of numbers measurements or observations, which are of interest.
- Size of the population N is the number of objects or observations in the population.
- Population may be finite or infinite.
- A finite sub-set of the population is known as **Sample.** Size of the sample is denoted by n.
- **Sampling** is the process of drawing the samples from a given population.
- If $n \geq 30$ the sampling is said to be **large sampling.**
- If $n < 30$ then the sampling is said to be **Small sampling.**
- **Statistical inference** deals with the methods of arriving at valid generalizations and predictions about the population using the information contained in the sample.
- **Parameters** Statistical measures or constants obtained from the population are known as population parameters or simply parameters.
- Population f(x) is a population whose probability distribution is f(x).If f(x) is binomial, Poisson or normal then the corresponding population is known as Binomial Population, Poisson population or normal Population.
- Samples must be representative of the population, sampling should be random.
- Random Sampling is one in which each member of the population has equal chances or probability of being included in the sample.
- Sampling where each member of a population may be chosen, more than once is called **Sampling with replacement**. A finite population, which is sampled with replacement, can theoretically be considered infinite since samples of any size can be drawn with out exhausting the population. For most practical purpose sampling from a finite population, which is very large, can be considered as sampling from an infinite population.
- If each member cannot be chosen more than once it is called sampling with out replacement.
- Any quantity obtained from a sample for the purpose of estimating a population parameter is called a sample statistics or briefly Statistic. Mathematically a sample statistic for a sample of size n can be defined as a function of the random variables X_1, $X_2 \ldots\ldots X_n$ i.e., $g(X_1, X_2 \ldots\ldots X_n)$. The function $g(X_1, X_2 \ldots\ldots X_n)$ is another random variable whose values can be represented by $g(X_1, X_2 \ldots\ldots X_n)$. The word statistic is often used for the r.v. or for its values.
- Random samples (Finite population): A set of observations $X_1, X_2 \ldots\ldots X_n$, constitute a random sample of size n from a finite population of size N, if its values are chosen so that

each subset of n of the N elements of the population has same probability if being selected.

- Random sample (Infinite Population): A set of observations X_1, X_2……X_n constitute a random sample of size n from infinite population f(x) if:

  (i) Each X_i is a r.v. whose distribution is given by f(x)

  (ii) These n r.v.'s are independent

- **Sample Mean** X_1, X_2……X_n is a random sample of size n the sample mean is a r.v. defined by $\overline{X} = \frac{X_1 + X_2 + ...X_n}{n}$
- **Sample Variance** X_1, X_2……X_n is a random sample of size n the sample variance is a r.v. defined by $S^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n}$ and is a measure of variability of data about the mean.
- **Sample Standard deviation** is the positive square root of the sample variance.
- **Degrees of freedom** (**d o f**) of a statistic is the positive integer denoted by υ, equals to n-k where n is the number of independent observations of the random sample and k is the number of population parameters which are calculated using sample data. Thus **d o f** υ = n – k is the difference between n, the sample size and k, the number of independent constraints imposed on the observations in the sample.
- **Sampling Distributions:** The probability distribution of a sample statistic is often called as sampling distribution of the statistic.
- The standard deviation of the sampling distribution of a statistic is called **Standard Error(S.E)**
- The mean of the sampling distribution of means, denoted by $\mu_{\overline{x}}$, is given by E($\overline{X}$) = $\mu_{\overline{x}}$ = μ where μ is the mean of the population.
- If a population is infinite or if sampling is with replacement, then the variance of the sampling distribution of means, denoted by $\sigma_{\overline{x}}^{2}$ is given by $E[(\overline{X} - \mu)^2] = \sigma_{\overline{x}}^{2} = \frac{\sigma^2}{n}$ where σ^2 is the variance of the population.
- If the population is of siqe N, if sampling is without replacement, and if the sample size is n≤N then $\sigma_{\overline{x}}^{2} = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)$
- The factor $\left(\frac{N-n}{N-1}\right)$ is called the finite population correction factor, is close to 1 (and can be omitted for most practical purposes) unless the samples constitutes a substantial portion of the population.
- (Central limit theorem) If $\overline{X}$ is the mean of a sample of size n taken from a population having the mean μ and the finite variance σ^2 , then $Z = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$ is a r.v. whose distribution function approaches that of the standard normal distribution as n→∞

- If $\overline{X}$ is the mean of a sample of size n taken from a finite population of size N with mean μ and variance σ^2 then Z= $\dfrac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}}$ is a r.v whose distribution function approaches that of the standard normal distribution as $n\to\infty$
- The normal distribution provides an excellent approximation to the sampling distribution of the mean $\overline{X}$ for n as small as 25 or 30
- If the random samples come from a normal population, the sampling distribution of the mean is normal regardless of the size of the sample

Inferences concerning means

- Statistical decisions are decisions or conclusions about the population parameters on the basis of a random sample from the population.
- Statistical hypothesis is an assumption or conjecture or guess about the parameters of the population distribution
- **Null Hypothesis (N.H)** denoted by H_0 is statistical hypothesis, which is to be actually tested for acceptance or rejection. NH is the hypothesis, which is tested for possible rejection under the assumption that it is true.
- Any Hypothesis which is complimentary to the N.H is called an **Alternative Hypothesis** denoted by H_1
- Simple Hypothesis is a statistical Hypothesis which completely specifies an exact parameter. N.H is always simple hypothesis stated as a equality specifying an exact value of the parameter. E.g. N.H = $H_0 : \mu = \mu_0$ N.H. = $H_0:\mu_1 - \mu_2 = \delta$
- Composite Hypothesis is stated in terms of several possible values.
- Alternative Hypothesis(A.H) is a composite hypothesis involving statements expressed as inequalities such as < , > or ≠

  i) A.H : $H_1:\mu>\mu_0$ (Right tailed) ii) A.H : $H_1:\mu<\mu_0$ (Left tailed)

  iii) A.H : $H_1:\mu\neq\mu_0$ (Two tailed alternative)

- **Errors in sampling**

  Type I error: Reject H_0 when it is true

  Type II error: Accept $\mathbf{H_0}$ when it is wrong (i.e) accept if when H_1 is true.

	Accept H_0	Reject H_0
H_0 is True	Correct Decision	Type 1 error
H_0 is False	Type 2 error	Correct Decision

- If P{ Reject H_0 when it is true}= P{ Reject H_0 | H_0}=α and

P{ Accept H_0 when it is false}= P{ Accept $H_0 \mid H_1$} = β then α,β are called the sizes of Type I error and Type II error respectively. In practice, type I error amounts to rejecting a lot when it is good and type II error may be regarded as accepting the lot when it is bad.

α and β are referred to as producers risk and consumers risk respectively.

- A region (corresponding to a statistic t) in the sample space S that amounts to rejection of H_0 is called critical region of rejection.
- Level of significance is the size of the type I error (or maximum producer's risk)
- The levels of significance usually employed in testing of hypothesis are 5% and 1% and is always fixed in advance before collecting the test information.
- A test of any statistical hypothesis where AH is one tailed(right tailed or left tailed) is called a **one–tailed test.** If AH is two-tailed such as: H_0**:**$\mu = \mu_0$, against the AH. H_1**:**$\mu \neq \mu_0$ ($\mu > \mu_0$ and $\mu < \mu_0$) is called **Two-Tailed Test.**
- The value of test statistics which separates the critical (or rejection) region and the acceptance region is called **Critical value or Significant value**. It depends upon (i) The level of significance used and (ii) The Alternative Hypothesis, whether it is two-tailed or single tailed
- From the normal probability tables we get

Critical Value (Z_α)	Level of significance (α) 1%	 5%	 10%
Two-Tailed test	$-Z_{\alpha/2} = -2.58$ $Z_{\alpha/2} = 2.58$	$-Z_{\alpha/2} = -1.96$ $Z_{\alpha/2} = 1.96$	$-Z_{\alpha/2} = -1.645$ $Z_{\alpha/2} = 1.645$
Right-Tailed test	$Z_\alpha = 2.33$	$Z_\alpha = 1.645$	$Z_\alpha = 1.28$
Left-Tailed Test	$-Z_\alpha = -2.33$	$- Z_\alpha = -1.645$	$-Z_\alpha = -1.28$

- When the size of the sample is increased, the probability of committing both types of error I and II (i.e) α and β are small, the test procedure is good one giving good chance of making the correct decision.
- P-value is the lowest level (of significance) at which observed value of the test statistic is significant.
- A test of Hypothesis (T. O.H) consists of
  1. Null Hypothesis (NH) : H_0
  2. Alternative Hypothesis (AH) : H_1
  3. Level of significance: α
  4. Critical Region pre determined by α
  5. Calculation of test statistic based on the sample data.
  6. Decision to reject NH or to accept it.

PROBLEMS:1. A population consists of five numbers 2,3,6,8 and 11. Consider all possible samples of size two which can be drawn with replacement from this population. Find

The mean of the population
The standard deviation of the population
The mean of the sampling distribution of means
The standard deviation of the sampling distribution of means

Solution: Given that N=5, n=2 and

i. Mean of the population

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

ii. Variance of the population

$$\sigma^2 = \sum \frac{(x_i - \bar{x})^2}{N} = \frac{(2-6)^2 + (3-6)^2 + (6-6)^2 + (8-6)^2 + (11-6)^2}{5}$$

$$= \frac{16+9+0+4+25}{5}$$

$$= 10.8$$

$$\sigma = 3.29$$

Sampling with replacement(infinite population):
The total number of samples with replacement is

$$N^n = 5^2 = 25$$

There 25 samples can be drawn

$$\left\{\begin{matrix} (2,2) & (2,3) & (2,6) & (2,8) & (2.11) \\ (3,2) & (3,3) & (3,6) & (3,8) & (3,11) \\ (6,2) & (6,3) & (6.6) & (6,8) & (6,11) \\ (8,2) & (8,3) & (8,6) & (8,8) & (8,11) \\ (11,2) & `(11,3) & (11,6) & (11,8) & (11,11) \end{matrix}\right\}$$

The sample means are

$$\left\{\begin{matrix} 2 & 2.5 & 4 & 5 & 6.5 \\ 2.5 & 3 & 4.5 & 5.5 & 7 \\ 4 & 4.5 & 6 & 7.0 & 8.5 \\ 5 & 5.5 & 7 & 8 & 9.5 \\ 6.5 & 7 & 8.5 & 9.5 & 11 \end{matrix}\right\}$$

iii. The mean of the sampling distribution of means is

$$\mu_{\bar{x}} = \frac{2+2.5+4+5+6.5+----+11}{25}$$

=6

iv. The standard deviation of the sampling distribution of means

$$\sigma_{\bar{x}}^2 = \frac{(2-6)^2 + (2.5-6)^2 + ----+(11-6)^2}{25}$$

$$=5.40$$

$$\sigma_{\bar{x}} = 2.32$$

2. **A population consists of five numbers4, 8, 12, 16, 20, 24. Consider all possible samples of size two which can be drawn without replacement from this population. Find**

i) The mean of the population
ii) The standard deviation of the population
iii)The mean of the sampling distribution of means
iv) the standard deviation of the sampling distribution of means

Solution: Given that N=6, n=2 and

i. Mean of the population

$$\mu = \sum \frac{x_i}{N} = \frac{4+8+12+16+20+24}{6} = \frac{84}{6} = 14$$

ii. Variance of the population

$$\sigma^2 = \sum \frac{(x_i - \bar{x})^2}{N} = \frac{(4-14)^2 + (8-14)^2 + (12-14)^2 + (16-14)^2 + (20-14)^2 + (24-14)^2}{6}$$

$$= \frac{100+36+4+4+36+100}{6}$$

$$= 46.67$$

$$\sigma = 3.29$$

Sampling without replacement (finite population):

The total number of samples without replacement is$N_{c_n} = 6_{c_2} = 15$

There 15 samples can be drawn

$$\left\{\begin{array}{lllll} (4,8) & (4,12) & (4,16) & (4,20) & (4,24) \\ (8,12) & (8,16) & (8,20) & (8,24) & \\ (12,16) & (12,20) & (12,24) & & \\ (16,20) & (16,24) & & & \\ (20,24) & \text{`} & & & \end{array}\right\}$$

The sample means are

$$\left\{\begin{array}{lllll} 6 & 8 & 10 & 12 & 14 \\ 10 & 12 & 14 & 16 & \\ 14 & 16 & 18 & & \\ 18 & 20 & & & \\ 22 & & & & \end{array}\right\}$$

iii. The mean of the sampling distribution of means is

$$\mu_{\bar{x}} = \frac{6+8+10+12-----20+22}{15}$$

=14

iv. The standard deviation of the sampling distribution of means

$$\sigma_{\bar{x}}^2 = \frac{(6-14)^2+(8-14)^2+----+(22-14)^2}{15}$$

=18.67

$$\sigma_{\bar{x}} = 4.32$$

3. 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.

Solution: Given mean of the population (μ) = 155 cm

Standard deviation of the population (σ) = 15 cm

Sample size (n) = 36

Mean of sample ($\bar{x}$) = 157 cm

Now $Z = \dfrac{\bar{x} - \mu}{\dfrac{\sigma}{\sqrt{n}}}$

$= \dfrac{157 - 155}{\dfrac{15}{\sqrt{36}}}$

=0.8

$\therefore P(\bar{x} \le 157) = P(Z < 0.8)$

$= 0.5 + P(0 \le Z \le 0.8)$

=0.5+0.2881

$\therefore P(\bar{x} \le 157)$ =0.7881

Exercise Problems:

1. Samples of size 2 are taken from the population 1, 2, 3, 4, 5, 6. Which can be drawn without replacement? Find

i) The mean of the population

ii) The standard deviation of the population

iii) The mean of the sampling distribution of means

iv) The standard deviation of the sampling distribution of means

2. If a 1-gallon can of paint covers on an average 513 square feet with a standard deviation of 31.5 square feet, what is the probability that the mean area covered by a sample of 40 of these 1-gallon cans will be anywhere from 510to 520 square feet?

- Test statistic for T.O.H. in several cases are

1. Statistic for test concerning mean σ known
   Z = $\dfrac{\overline{X} - \mu_0}{\sigma / \sqrt{n}}$

2. Statistic for large sample test concerning mean with σ unknown

$$Z = \frac{\overline{X} - \mu_0}{S/\sqrt{n}}$$

3. Statistic for test concerning difference between the means

$$Z = \frac{\left(\overline{X_1} - \overline{X}_2\right) - \delta}{\sqrt{\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)}}$$ under NH $H_0: \mu_1 - \mu_2 = \delta$ against the AH, $H_1: \mu_1 - \mu_2 > \delta$ or $H_1: \mu_1 - \mu_2 < \delta$ or $H_1: \mu_1 - \mu_2 \neq \delta$

4. Statistic for large samples concerning the difference between two means (σ_1 and σ_2 are unknown)

$$Z = \frac{\left(\overline{X_1} - \overline{X}_2\right) - \delta}{\sqrt{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)}}$$

Statistics for large sample test concerning one proportion

$$Z = \frac{X - np_o}{\sqrt{np_0(1-p_0)}}$$ under the N.H: $H_0: p = p_o$ against $H_1: p \neq p_0$ or $p > p_0$ or $p < P_0$

Statistic for test concerning the difference between two proportions

$$Z = \frac{\frac{X_1}{n_1} - \frac{X_2}{n_2}}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}}$$ with $\hat{p} = \frac{X_1 + X_2}{n_1 + n_2}$ under the NH : $H_0: p_1 = p_2$ against the AH $H_1: p_1 < p_2$ or $p_1 > p_2$ or $p_1 \neq p_2$

- To determine if a population follows a specified known theoretical distribution such as ND,BD,PD the χ^2(chi-square) test is used to assertion how closely the actual distribution approximate the assumed theoretical distribution. This test is based on how good a fit is there between the observed frequencies and the expected frequencies is known as "**goodness-of-fit-test**".
- Large sample confidence interval for p

$$\frac{x}{n} - Z_{\alpha/2}\sqrt{\frac{\frac{x}{n}\left(1 - \frac{x}{n}\right)}{n}} < p < \frac{x}{n} + Z_{\alpha/2}\sqrt{\frac{\frac{x}{n}\left(1 - \frac{x}{n}\right)}{n}}$$ where the degree of confidence is 1- α

- Large sample confidence interval for difference of two proportions (p_1- p_2) is

$$\left(\frac{x_1}{n_1}-\frac{x_2}{n_2}\right)\pm Z_{\alpha/2}\sqrt{\frac{\frac{x_1}{n_1}\left(1-\frac{x_1}{n_1}\right)}{n_1}+\frac{\frac{x_2}{n_2}\left(1-\frac{x_2}{n_2}\right)}{n_2}}$$

- Maximum error of estimate E = $Z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$ with observed value x/n substituted for p we obtain an estimate of E
- Sample size n = p(1-p)$\left(\frac{Z_{\alpha/2}}{E}\right)^2$ when p is known

  n= $\frac{1}{4}\left(\frac{Z_{\alpha/2}}{E}\right)^2$ when p is unknown

- One sided confidence interval is of the form p < (1/2n)χ_α^2 with (2n+1) degrees of freedom.

Problems:

1. **A sample of 400 items is taken from a population whose standard deviation is 10.The mean of sample is 40.Test whether the sample has come from a population with mean 38 also calculate 95% confidence interval for the population.**

   Solution: Given n=400, $\bar{x}=40$ and μ =38 and σ =10

   1. **Null hypothesis(H$_0$):** μ =38
   2. **Alternative hypothesis(H$_1$):** $\mu \neq 38$
   3. **Level of significance:** α =0.05 and Z_α =1.96
   4. **Test statistic:** $Z=\frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}}$

      $Z=\frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}}=\frac{40-38}{\frac{10}{\sqrt{400}}}$ =4

      $|Z|=4$
   5. **Conclusion:**

      $\therefore |Z|>Z_\alpha$

      $\therefore$ We reject the Null hypothesis.

$$\text{Confidence interval} = \left(\bar{x} - Z_{\alpha} \frac{\sigma}{\sqrt{n}}, \bar{x} + Z_{\alpha} \frac{\sigma}{\sqrt{n}} \right)$$

$$= \left(40 - 1.96 \frac{10}{\sqrt{400}}, 40 + 1.96 \frac{10}{\sqrt{400}} \right)$$

$$= (39.02, 40.98)$$

2. Samples of students were drawn from two universities and from their weights in kilograms mean and S.D are calculated and shown below make a large sample test to the significance of difference between means.

	MEAN	S.D	SAMPLE SIZE
University-A	55	10	400
University-B	57	15	100

Solution: Given n$_1$=400, n$_2$=100, $\bar{x}_1$=55, $\bar{x}_2$ =57
S$_1$=10 and S$_2$=15

1. **Null hypothesis(H$_0$):** $\bar{x}_1 = \bar{x}_2$
2. **Alternative hypothesis(H$_1$):** $\bar{x}_1 \neq \bar{x}_2$
3. **Level of significance:** α =0.05 and Z_{α} =1.96
4. **Test statistic:** $Z = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}}} = \dfrac{55 - 57}{\sqrt{\dfrac{100}{400} + \dfrac{225}{100}}}$ =-1.26

$$|Z| = 1.26$$

5. **Conclusion:**

$\therefore |Z| < Z_{\alpha}$

$\therefore$ We accept the Null hypothesis.

3. **In a sample of 1000 people in Karnataka 540 are rice eaters and the rest are wheat eaters. Can we assume that both rice and wheat are equally popular in this state at 1% level of significance?**

Solution: Given n =400, x =540

$$p = \frac{x}{n} = \frac{540}{1000} = 0.54$$

$$P = \frac{1}{2} = 0.5\ ,\ Q = 0.5$$

1. **Null hypothesis(H_0):** P =0.5
2. **Alternative hypothesis(H_1):** $P \neq 0.5$
3. **Level of significance:** α **=**1% and Z_α =2.58
4. **Test statistic:** $Z = \frac{P-p}{\sqrt{\frac{PQ}{n}}}$

$$Z = \frac{P-p}{\sqrt{\frac{PQ}{n}}} = \frac{0.54-0.5}{\sqrt{\frac{0.5\times 0.5}{1000}}} = 2.532$$

$$|Z| = 2.532$$

5. **Conclusion:**

$\therefore |Z| < Z_\alpha$

$\therefore$ We accept the Null hypothesis.

4. **Random sample of 400 men and 600 women were asked whether they would like to have flyover near their residence .200 men and 325 women were in favour of proposal. Test the hypothesis that the proportion of men and women in favour of proposal are same at 5% level.**

Solution: Given n_1=400, n_2=600 , $x_1 = 200$ and $x_2 = 325$

$$p_1 = \frac{200}{400} = 0.5$$

$$p_2 = \frac{325}{600} = 0.541$$

$$p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{400\times\frac{200}{400} + 600\times\frac{325}{600}}{400+600} = 0.525$$

$q = 1 - p = 1 - 0.525 = 0.475$

1. **Null hypothesis(H$_0$):** $p_1 = p_2$
2. **Alternative hypothesis(H$_1$):** $p_1 \neq p_2$
3. **Level of significance:** α =0.05 and Z_α =1.96
4. **Test statistic:** $Z = \frac{p_1 - p_2}{\sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$ =

$$\frac{0.5 - 0.541}{\sqrt{0.525 \times 0.425\left(\frac{1}{400} + \frac{1}{600}\right)}} = -1.28$$

$$|Z| = 1.28$$

5. **Conclusion:**

$\therefore |Z| < Z_\alpha$

$\therefore$ We accept the Null hypothesis.

Exercise Problems:

1. An ambulance service claims that it takes on the average 8.9 minutes to reach its destination In emergency calls. To check on this claim the agency which issues license to Ambulance service has then timed on fifty emergency calls getting a mean of 9.2 minutes with 1.6 minutes. What can they conclude at 5% level of significance?

2.According to norms established for a mechanical aptitude test persons who are 18 years have an average weight of 73.2 with S.D 8.6 if 40 randomly selected persons have average 76.7 test the hypothesis H_0 : μ =73.2 againist alternative hypothesis : μ>73.2.

3.A cigarette manufacturing firm claims that brand A line of cigarettes outsells its brand B by 8% .if it is found that 42 out of a sample 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.

4. The nicotine in milligrams of two samples of tobacco were found to be as follows. Test the hypothesis for the difference between means at 0.05 level

Sample-A	24	27	26	23	25	
Sample-B	29	30	30	31	24	36

5. A machine puts out of 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 improved?

MODULE – V

TEST OF HYPOTHESIS – II

- Maximum error E of estimate of a normal population mean μ with σ unknown by using small sample

  mean $\overline{X}$ is E = $t_{\alpha/2}\frac{S}{\sqrt{n}}$ sample size n = $\left[t_{\alpha/2}\frac{S}{E}\right]^2$ here the percentage of confidence is (1 - α)100% and the degree of confidence is 1-α

- Small sample confidence interval for μ

  $\bar{x} - t_{\alpha/2}\frac{S}{\sqrt{n}} < \mu < \bar{x} + t_{\alpha/2}\frac{S}{\sqrt{n}}$

- If $\overline{X}$ is the mean of a random sample of size n taken from a normal population having the mean μ and

  the variance σ^2 , and $S^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}$ then t= $\frac{\overline{X}-\mu}{\frac{s}{\sqrt{n}}}$ is a r.v. having the

  t- distribution with the parameter υ = (n-1)dof

- The overall shape of a t-distribution is similar to that of a normal distribution both are bell shaped and symmetrical about the mean. Like the standard normal distribution t-distribution has the mean 0, but its variance depends on the parameter υ (nu), called the number of degrees of freedom. The variance of t- distribution exceeds1, but it approaches 1 as n→∞. The t-distribution with υ-degree of freedom approaches the standard normal distribution as υ→∞.

- The standard normal distribution provides a good approximation to the t- distribution for samples of size 30 or more.

- If S^2 is the variance of a random sample of size n taken from a normal population having the variance σ^2 , then $\chi^2 = \frac{(n-1)S^2}{\sigma^2} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{\sigma^2}$ is a r.v. having the chi-square distribution with the parameter υ = n-1

- The chi-square distribution is not symmetrical
- If S_1^2 and S_2^2 are the variances of independent random samples of size n_1 and n_2 respectively, taken from two normal populations having the same variance, then $F = \frac{S_1^2}{S_2^2}$ is a r.v. having the F- distribution with the parameter's υ_1=n_1-1 and υ_2=n_2-1 are called the numerator and denominator degrees of freedom respectively.
- $F_{1-\alpha}(\upsilon_1,\upsilon_2) = \frac{1}{F_\alpha(\upsilon_2,\upsilon_1)}$

Problems:

1. **Producer of 'gutkha' claims that the nicotine content in his 'gutkha' on the average is 83 mg. can this claim be accepted if a random sample of 8 'gutkhas' of this type have the nicotine contents of 2.0,1.7,2.1,1.9,2.2,2.1,2.0,1.6 mg.**

Solution: Given n=8 and μ =1.83 mg

6. **Null hypothesis(H$_0$):** μ =1.83
7. **Alternative hypothesis(H$_1$):** $\mu \neq 1.83$
8. **Level of significance:** α =0.05

   t_α for n-1 degrees of freedom

   $t_{0.05}$ for 8-1 degrees of freedom is 1.895
9. **Test statistic:** $t = \frac{\bar{x} - \mu}{\frac{S}{\sqrt{n}}}$

x	$(x - \bar{x})$	$(x - \bar{x})^2$
2.0	0.05	0.0025
1.7	-0.25	0.0625
2.1	0.15	0.0225
1.9	-0.05	0.0025
2.2	0.25	0.0625

2.1	0.15	0.0225
2.0	0.05	0.0025
1.6	-0.35	0.1225
Total=15.6		

$$\bar{x} = \frac{15.6}{8} = 1.95 \text{ and } S^2 = \sum \frac{(x-\bar{x})^2}{n-1} = \frac{0.3}{7}$$

S=0.21

$$t = \frac{\bar{x}-\mu}{\frac{S}{\sqrt{n}}} = \frac{1.95-1.83}{\frac{0.21}{\sqrt{8}}} = 1.62$$

$$|t| = 1.62$$

10. Conclusion:

$\therefore |t| < t_\alpha$

$\therefore$ We accept the Null hypothesis.

2. **The means of two random samples of sizes 9,7 are 196.42 and 198.82.the sum of squares of deviations from their respective means are 26.94,18.73.can the samples be considered to have been the same population?**

Solution: Given n$_1$=9, n$_2$=7, $\bar{x}_1$=196.42, $\bar{x}_2$=198.82 and $\sum(x_i-\bar{x}_1)^2$=26.94, $\sum(x_i-\bar{x}_2)^2$=18.73

$$\therefore S^2 = \frac{\sum(x_i-x_1)^2+\sum(x_i-x_2)^2}{n_1+n_2-2} = 3.26$$

$\Rightarrow$ S=1.81

Null hypothesis(H$_0$): $\bar{x}_1 = \bar{x}_2$

Alternative hypothesis(H$_1$): $\bar{x}_1 \neq \bar{x}_2$

Level of significance: α =0.05

t_α for n_1+n_2-2 degrees of freedom

$t_{0.05}$ for 9+7-2=14 degrees of freedom is 2.15

Test statistic: $t = \dfrac{\bar{x}_1 - \bar{x}_2}{S\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$ = $\dfrac{196.42 - 198.82}{(1.81)\sqrt{\dfrac{1}{9} + \dfrac{1}{7}}}$ =-2.63

$|t| = 2.63$

Conclusion:

$\therefore |t| > t_\alpha$

$\therefore$ We reject the Null hypothesis.

3. **In one sample of 8 observations the sum of squares of deviations of the sample values from the sample mean was 84.4 and another sample of 10 observations it was 102.6 .test whether there is any significant difference between two sample variances at at 5% level of significance.**

Solution: Given n$_1$=8, n$_2$=10, $\sum(x_i - \bar{x}_1)^2$ =84.4 and $\sum(x_i - \bar{x}_2)^2$ =102.6

$$S_1{}^2 = \frac{\sum(x_i - x_1)^2}{n_1 - 1} = \frac{84.4}{7} = 12.057$$

$$S_2{}^2 = \frac{\sum(x_i - x_1)^2}{n_2 - 1} = \frac{102.6}{9} = 11.4$$

1. **Null hypothesis(H$_0$):** $S_1{}^2 = S_2{}^2$
2. **Alternative hypothesis(H$_1$):** $S_1{}^2 \neq S_2{}^2$
3. **Level of significance:** α =0.05

   F_α For $(n_1 - 1, n_2 - 1)$ degrees of freedom

   $F_{0.05}$ For (7,9) degrees of freedom is 3.29
4. **Test statistic:** $F = \dfrac{S_1^2}{S_2^2}$ $= \dfrac{12.057}{11.4}$ =1.057

$$|F| = 1.057$$

5. **Conclusion:**

$\therefore |F| < F_\alpha$

$\therefore$ We accept the Null hypothesis.

4. **The following table gives the classification of 100 workers according to gender and nature of work. Test whether the nature of work is independent of the gender of the worker.**

	Stable	**Unstable**	**Total**
Male	**40**	**20**	**60**
Female	**10**	**30**	**40**
Total	**50**	**50**	**100**

Solution: Given that

Expected frequencies $= \frac{\text{row total} \times \text{column total}}{\text{grand total}}$

$\frac{90 \times 100}{200}$ =45	$\frac{90 \times 100}{200}$ =45	90
$\frac{90 \times 100}{200}$ =55	$\frac{90 \times 100}{200}$ =55	110
100	100	200

Calculation of χ^2 :

Observed Frequency(O_i)	Expected Frequency(E_i)	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
60	45	225	5
30	45	225	5
40	55	225	4.09
70	55	225	4.09
			18.18

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} = 18.18$$

1. **Null hypothesis(H_0):** $O_i = E_i$

2. **Alternative hypothesis(H_1):** $O_i \neq E_i$

3. **Level of significance:** α =0.05

   $\chi_{\alpha}{}^{2}$ For (r-1)(c-1) degrees of freedom

   $\chi_{0.05}{}^{2}$ For (2-1)(2-1)=1 degrees of freedom is 3.84

4. **Test statistic:** $\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$ =18.18

   $$\left|\chi^2\right| = 1.057$$

5. **Conclusion:**

   $\therefore \left|\chi^2\right| > \chi_{\alpha}{}^{2}$

   $\therefore$ We reject the Null hypothesis.

Exercise problems:

1. Two random samples gave the following results

Sample	size	Sample mean	Sum of squares of deviations from mean
I	10	15	90
II	12	14	108

Test whether the samples came from the same population or not?

200 digits were chosen at random from set of tables the frequency of the digits are

2. Use chi square test to asset the correctness of the hypothesis that the digits are distributed in equal number in the table

digit	0	1	2	3	4	5	**6**	**7**	**8**	**9**
frequency	18	19	23	21	16	25	22	20	21	15

3. 5 dice were thrown 96 times the number of times showing 4,5 or 6 obtain is given below Fit a binomial distribution and test for goodness of fit

Practise problems:

x	0	1	2	3	4	5
frequency	1	10	24	35	18	8

Part-A

1 Producer of ‘gutkha’ claims that the nicotine content in his ‘gutkha’ on the average is 0.83 mg. can this claim be accepted if a random sample of 8 ‘gutkhas’ of this type have the nicotine contents of 2.0,1.7,2.1, 1.9,2.2, 2.1, 2.0,1.6 mg.

2 A sample of 26 bulbs gives a mean life of 990 hrs with S.D of 20hrs. The manufacturer claims that the mean life of bulbs 1000 hrs. Is the sample not upto the standard?

3 A random sample of 10 boys had the following I.Q’s 70,120,110,101,88,83,95,98,107,100. Do the data support the assumption of population means I.Q of 100. Test at 5% level of significance?

4 The means of two random samples of sizes 9,7 are 196.42 and 198.82.the sum of squares of deviations from their respective means are 26.94,18.73.can the samples be considered to have been the same population?

5 In one sample of 8 observations the sum of squares of deviations of the sample values from the sample mean was 84.4 and another sample of 10 observations it was 102.6 .test whether there is any significant difference between two sample variances at at 5% level of significance.

6 Two random samples gave the following results.

Sample	**size**	**Sample mean**	**Sum of squares of deviations from mean**
I	10	15	90
II	12	14	108

Test whether the samples came from the same population or not?

7 Two independent samples of items are given respectively had the following values.

Sample I	11	11	13	11	15	9	12	14
Sample II	9	11	10	13	9	8	10	-

Test whether there is any significant difference between their means?

8 Time taken by workers in performing a job by method 1 and method 2 is given below.

Method 1	20	16	27	23	22	26	-
Method 2	27	33	42	35	32	34	38

Does the data show that variances of time distribution from population which these samples are drawn do not differ significantly?

9 The no. of automobile accidents per week in a certain area as follows:

12,8,20,2,14,10,15,6,9,4. Are these frequencies in agreement with the belief that accidents were same in the during last 10 weeks.

10 A die is thrown 264 times with the following results .show that the die is unbiased.

No appeared on die	1	2	3	4	5	6
Frequency	40	32	28	58	54	52

11 200 digits were choosen at random from set of tables the frequency of the digits are

digit	0	1	2	3	4	5	6	7	8	9
frequency	18	19	23	21	16	25	22	20	21	15

Use chi square test to asset the correctness of the hypothesis that the digits are distributed in equal number in the table

12 Fit a poisson distribution to the following data and test the goodness of fit at 0.05 level.

x	0	1	2	3	4	5	6	7
frequency	305	366	210	80	28	9	2	1

13 Given below is the number of male births in 1000 families having 5 children

Male children	0	1	2	3	4	5
Number of families	40	300	250	200	30	180

Test whether the given data is consistent with the hypothesis that the binomial distribution holds if the chance of a male birth is equal to female birth.

14 5 dice were thrown 96 times the number of times showing 4,5 or 6 obtain is given below

x	0	1	2	3	4	5
frequency	1	10	24	35	18	8

Fit a binomial distribution and test for goodness of fit.

15 The following is the distribution of the hourly number of trucks arriving at a company

Trucks per hour	0	1	2	3	4	5	6	7	8
frequency	52	151	130	102	45	12	3	1	2

wear house.

Fit a poisson distribution to the following table and test the goodness of fit at 0.05 level.

16 The average breaking strength of the steel rods is specified to be 18.5 thousand pounds. To test this sample of 14 rods were tested. The mean and S.D obtained were 17.85 and 1.955 respectively. Is the result of experiment significant?

17 A group of 5 patients treated with medicine A weigh 42, 39, 48, 60 and 41 kgs . Second group of 7 patients from the same hospital treated with medicine B weigh 38, 42, 56, 64, 68, 69 and 62 kgs. Do you agree with the claim that medicine B increases the weigh significantly.

18 In one sample of 10 observations, the sum of the deviations of the sample values from sample mean was 120 and in the other sample of 12 observations it was 314. Test whether the difference is significant at 5% level.

19 The following table gives the classification of 100 workers according to gender and nature of work. Test whether the nature of work is independent of the gender of the worker.

	Stable	**Unstable**	**Total**
Male	40	20	60
Female	10	30	40
Total	50	50	100

20 The following random samples are measurements of the heat-producing capacity (in millions of calories per ton) of speciments of coal from two mines:

Mine 1	8,260	8,130	8,350	8,070	8,340	…
Mine 2	7,950	1,890	7,900	8,140	7,920	7,840

Use the 0.05 level of significance to test whether it is reasonable to assume that the variances of the two populations are equal.

Part B

1 A mechanist making engine parts with axle diameters of 0.700 inch. A random sample of 10 parts shows a mean diameter of 0.742 inch with a S.D of 0.040 inch. Compute the statistic you would use to test whether the work is meeting the specifications.

2 To examine the hypothesis that the husbands are more intelligent than the wives, an investigator took a sample of 10 couples and administered them a test which measures the I.Q. The results are as follows.

Husbands	117	105	97	105	123	109	86	78	103	107
Wives	106	98	87	104	116	95	90	69	108	85

Test the hypothesis with a reasonable test at the level of significance of 0.05.

3 Two independent samples of 8 & 7 items respectively had the following values.

Sample I	11	11	13	11	15	9	12	14
Sample II	9	11	10	13	9	8	10	

Is the difference between the means of samples significant?

4 Pumpkins were grown under two experimental conditions. Two random samples of 11 and 9 pumpkins. the sample standard deviation of their weights as 0.8 and 0.5 respectively. Assuming that the weight distributions are normal, test hypothesis that the true variances are equal.

5 From the following data, find whether there is any significant liking in the habit of taking soft drinks among the categories of employees.

Soft drinks	**Clerks**	**Teachers**	**officers**
Pepsi	10	25	65
Thumsup	15	30	65

Fanta	50	60	30

6 In an investigation on the machine performance, the following results are obtained.

	No.of units inspected	**No.of defective**
Machine1	375	17
Machine2	450	22

7 A survey of 240 families with 4 children each revealed the following distribution.

Male Births	4	3	2	1	0
No of families	10	55	105	58	12

Test whether the male and female births are equally popular.

8 Samples of students were drawn from two universities and from their weights in kilograms mean and S.D are calculated and shown below make a large sample test to the significance of difference between means.

	Mean	**Standard Deviation**	**Sample Size**
University A	55	10	10
University B	57	15	20

9 The measurements of the output of two units have given the following results. Assuming that both samples have been obtained from the normal populations at 10% significant level, test whether the two populations have the same variance.

Unit-A	14.1	10.1	14.7	13.7	14.0
Unit - B	14.0	14.5	13.7	12.7	14.1

10 The nicotine in milligrams of two samples of tobacco were found to be as follows. Test the hypothesis for the difference between means at 0.05 level.

Sample-A	24	27	26	23	25	-
Sample-B	29	30	30	31	24	36