



A T M E
College of Engineering



Probability, Distributions and Statistics

1BMATCS301



Course outcome

- CO1:** Apply the concepts of modular arithmetic, congruences, linear Diophantine equations, Euler's theorem, Fermat's theorem, Wilson's theorem, and the RSA cryptographic algorithm to solve engineering and computational problems.
- CO2:** Apply least squares curve fitting, correlation, regression analysis, and rank correlation techniques to analyse engineering data and establish relationships between variables.
- CO3:** Solve engineering problems involving discrete and continuous random variables using probability distributions, mathematical expectation, variance, and standard probability models such as Binomial, Poisson and Normal distributions.
- CO4:** Analyse joint probability distributions and Markov chains by determining expectations, covariance, correlation, transition probabilities, stationary distributions, and absorbing states for stochastic systems.
- CO5:** Apply statistical inference techniques including hypothesis testing, confidence intervals, large and small sample tests, Student's t-test, Chi-square test, and the Central Limit Theorem to draw valid conclusions from engineering data.



ATME

College of Engineering



Module-1: Modular Arithmetic

Module-2: Statistics

Module-3: Probability Distributions

Module-4: Joint probability distribution

Module-5: Statistical Inference



ATME

College of Engineering



Module-1: Modular Arithmetic

Divisibility

The set of integers consists of all positive integers, all negative integers and zero. It is denoted by I or $Z = \{\dots, -3, -2, -1, 0, 1, 2, 3, \dots\}$

$$= \{0, \pm 1, \pm 2, \pm 3, \pm 4, \dots\}$$

Consider 2 integers a and b where $a \neq 0$, a divides b

If there exists an integer k such that $b = ka$

Eg: 9 divides 54, there is an integer 6 such that $54 = 6 \times 9$

Division Algorithm

Given two integers a and b where $a > 0$, two unique integers q and r such that $b = qa + r$ where $0 \leq r < a$. This is known as division algorithm. q is called the quotient and r is called the remainder.

Note: In all divisions q may be positive or negative.

Greatest Common Divisor(G.C.D) or Highest Common Factor(H.C.F)

The G.C.D of two integers a and b (both of them are not zero) is a unique positive integer d such that d is the common divisor of both a and b .

$$d/a, d/b$$

every common divisor of a and b divides d

$$x/a \text{ and } x/b \Rightarrow x/d$$

The G.C.D of 2 numbers a and b is written as (a, b)

$$\text{i.e., } d = (a, b)$$

Euclid's Algorithm method to find the G.C.D of 2 numbers a and b and to express the G.C.D as $ax+by$

A method of finding the greatest common divisor of two numbers by dividing the larger by the smaller, the smaller by the remainder, the first remainder by the second remainder and so on until exact division is obtained. Hence the greatest common divisor is the exact divisor.



Problems

Find the G.C.D of 32 and 54 and express it in the form of $32x + 54y$

32)54(1

32

$$22 = 54 - 1(32)$$

22

22)32(1

22

$$10 = 32 - 1(22)$$

10

10)22(2

20

$$2 = 22 - 2(10)$$

2

2)10(5

10

$$0 = 10 - 5(2)$$

0



ATME

College of Engineering



The last non zero remainder is 2.

$$2 = 22 - 2(10)$$

$$2 = [54 - 1(32)] - 2[32 - 1(22)]$$

$$= 54 - 3(32) + 2(22)$$

$$= 54 - 3(32) + 2[54 - 1(32)]$$

$$= 3(54) - 5(32)$$

$$= 54(3) + 32(-5)$$

$$2 = 54(x) + 32(y) \text{ where } x=3 \text{ and } y=-5$$

Relatively Prime numbers:

Two numbers a and b are said to be relatively prime or coprime if and only if $(a, b) = 1$, i.e. the G.C.D of a and b is 1

Example: $(8, 15) = 1$

8 and 15 are relatively prime



ATME

College of Engineering



Congruence's:

Let m be a positive integer (>1). If a and b are any integers then a is said to be congruent to b modulo m if and only if $m/(a-b)$

' a is congruent to b modulo m is written as $a \equiv b(modm)$

Example:

$25 \equiv 3(mod11)$, $11/25-3$, i.e $11/22$

$-69 \equiv -5(mod16)$, $16/-69+5$, i.e $16/-64$

Consider the congruence $134 \equiv 108(mod13)$ which is true

When 134 and 108 are divided by 13, the same remainder 4 is obtained. This gives the alternative definition of the congruence.

Definition of Congruence's:

If m is a positive integer (>1) and a and b are any integer the a is said to be congruent to b modulo m if and only if a and b leave the same remainder when divided by m

Note: 1) $a \equiv b(modm) \implies m/(a-b)$

Conversely $m/(a-b)$ then $a \equiv b(modm)$

Problems

Solve $5x \equiv 4 \pmod{13}$

$$\Rightarrow 13 \mid (5x - 4)$$

$$\Rightarrow 5x - 4 = 13k \text{ where } k \in \mathbb{Z}$$

$$\Rightarrow 5x = 13k + 4$$

$$\Rightarrow x = \frac{13k+4}{5}$$

By inspection, $k = 2$ gives the integral value of Z

i.e., $x=6$

$$x \equiv 6 \pmod{13}$$

The general solution is $x = 6 + 13t$ where $t \in \mathbb{Z}$



ATME

College of Engineering



Solve $78 + x \equiv 3 \pmod{5}$

$$78 + x - 3 \equiv 5n$$

$$75 + x \equiv 5n$$

Let $x = 5$

$$75 + 5 \equiv 80 \text{ (80 is multiple of 5)}$$

The least value of x is 5

Find the unit digit in the number 7^{289}

$$7^2 \equiv 49 \equiv 9 \pmod{10}$$

$$7^4 \equiv 2401 \equiv 1 \pmod{10}$$

$$(7^4)^{72} \equiv 1 \pmod{10}$$

$$7^{288} \equiv 1 \pmod{10} \text{-----(1)} , 7 \equiv 7 \pmod{10} \text{-----(2)}$$

From (1) & (2)

$$7^{289} \equiv 7 \pmod{10}$$

The unit digit in 7^{289} is 7

What is the remainder in the division of 2^{50} is divided by 7

$$2^3 \equiv 8 \equiv 1(\text{mod}7)$$

$$(2^3)^{16} \equiv 1(\text{mod}7) \text{ -----(1)}$$

$$2^2 \equiv 4(\text{mod}7) \text{ -----(2)}$$

From (1) & (2)

$$2^{50} \equiv 4(\text{mod}7)$$

The remainder when 2^{50} is divided by 7 is 4



ATME

College of Engineering



Find the remainder when $135 \times 74 \times 48$ is divided by 7.

7)135(1

7

65

63

2

$$135 \equiv 2 \pmod{7}$$

7)74(10

70

4

$$74 \equiv 4 \pmod{7}$$

7)48(6

42

6

$$48 \equiv 6 \pmod{7}$$

$$135 \times 74 \times 48 \equiv 48 \pmod{7} \equiv 6 \pmod{7}$$

Remainder is 6.

Rules for finding x in linear congruence : General format : $ax \equiv b(mod\ n)$

Find $GCD(a, n) = d$ (let)

If b/d is a whole number, then solution exists otherwise no solution

Find $d \mod n$ ---- these number of solutions are possible

Divide both sides by d

Multiply both sides by an inverse

$$(aa^{-1})x \equiv b.a^{-1}(mod\ n)$$

General solution is given by $x_k = x_0 + k(\frac{n}{d})$ where $k= 0,1,2,3.....$

Solve $14x \equiv 12 \pmod{18}$

$ax \equiv b \pmod{n}$

$a = 14, b = 12, n = 18$

1) $\text{GCD}(14, 18) = 2 = d$

2) $\frac{b}{d} = \frac{12}{2} = 6 = \text{solutions exists}$

3) $d \pmod{n} = 2 \pmod{18} = 2$ solutions exists

4) Divide both side by d

$(14x/2) \equiv (12/2) \pmod{18/2}$

$7x \equiv 6 \pmod{9}$

5) multiply both sides by 7^{-1}

$(77^{-1})x \equiv 6 \cdot 7^{-1} \pmod{9}$

$(7xc) \pmod{9} = 1$

$c=1, 7 \pmod{9} \neq 1$

$c=2, 14 \pmod{9} \neq 1$

$c=3, 21 \pmod{9} \neq 1$

$c=4, 28 \pmod{9} = 1$



ATME

College of Engineering



$$x \equiv 6 \cdot 7^{-1}(\text{mod } 9)$$

$$x \equiv 6 \cdot 4(\text{mod } 9)$$

$$x \equiv 24(\text{mod } 9)$$

$$x = 6$$

The general solution is given by $x_k = x_0 + k\left(\frac{n}{d}\right)$

$$x_1 = 6 + 1\left(\frac{18}{2}\right) = 15$$

THE CHINESE REMAINDER THEOREM

The Chinese Remainder theorem is used to solve a set of different congruent equations with one variable but different moduli which are relatively prime as shown below:

$$\begin{aligned}
&X \equiv a_1(mod\ m_1)\ , \\
&X \equiv a_2(mod\ m_2)\ , \\
&X \equiv a_3(mod\ m_3)\ , \\
&..... \\
&X \equiv a_n(mod\ m_n)\ ,
\end{aligned}$$

CRT states that the above equations have a unique solutions of the moduli are relatively prime

$$X = (a_1M_1M_1^{-1} + a_2M_2M_2^{-1} + a_3M_3M_3^{-1} + \cdots + a_nM_nM_n^{-1})mod\ M$$

- Solve the following equations using CRT

$$X \equiv 2(mod 3) \quad , \quad X \equiv 3(mod 5) , X \equiv 2(mod 7)$$

- G.C.D (3,5,7) = 1

$$X = (a_1M_1M_1^{-1} + a_2M_2M_2^{-1} + a_3M_3M_3^{-1})mod M$$

- $a_1 = 2, m_1 = 3, M_1 = 35, M_1^{-1} = 2$
- $a_2 = 3, m_2 = 5, M_2 = 21, M_2^{-1} = 1$
- $a_3 = 2, m_3 = 7, M_3 = 15, M_3^{-1} = 1$



ATME

College of Engineering



- $M = m_1 \times m_2 \times m_3 = 3 \times 5 \times 7 = 105$
- $M_1 = \frac{M}{m_1} = \frac{105}{3} = 35$; $M_2 = \frac{M}{m_2} = \frac{105}{5} = 21$; $M_3 = \frac{M}{m_3} = \frac{105}{7} = 15$
- $M_1 M_1^{-1} = 1(\text{mod } m_1)$ $M_2 M_2^{-1} = 1(\text{mod } m_2)$ $M_3 M_3^{-1} = 1(\text{mod } m_3)$
- $35 M_1^{-1} = 1 \text{mod } 3$ $21 M_2^{-1} = 1(\text{mod } 5)$ $15 M_3^{-1} = 1(\text{mod } 7)$
- $35 \times 2 = 1 \text{mod } 3$ $21 \times 1 = 1(\text{mod } 5)$ $15 \times 1 = 1(\text{mod } 7)$
- $M_1^{-1} = 2$ $M_2^{-1} = 1$ $M_3^{-1} = 1$
- $X = (a_1 M_1 M_1^{-1} + a_2 M_2 M_2^{-1} + a_3 M_3 M_3^{-1}) \text{mod } M$
- $X = (2.35.2 + 3.21.1 + 2.15.1) \text{mod } 105$
 $= 233 \text{ mod } 105$
 $= 23$

- **Linear Diophantine Equations:**

An equation of the form $ax + by + c = 0$ where $a \neq 0, b \neq 0$ and c is an integer is called a linear Diophantine equation in two variables x and y

Examples : $8x + 17y = 7, 2x + 3y = 12$

- **Solution of Linear Diophantine Equations:**

A pair (x, y) of integers is called a solution of Linear Diophantine Equations $ax + by = c$ if $ax_0 + by_0 = c$ and $\text{GCD}(a, b) = d$ iff d/c then solution exists, then general solution is given by

$$x_1 = x_0 + \frac{b}{d} t ; y_1 = y_0 - \frac{a}{d} t$$



ATME

College of Engineering



1) Find the general solutions of the equations $70x + 112y = 168$

$$G.C.D(70, 112) = 14$$

$$\text{Now } 14/168 = 1/12$$

Linear Diophantine Equations $70x + 112y = 168$ has a solution

By reverse

$$14 = 42 - 28$$

$$= 42 - (70 - 42)$$

$$= 42 - 70 + 42$$

$$= 2(42) - 70$$

$$= 2(112 - 70) - 70$$

$$= 2(112) - 2(70) - 70$$

$$14 = 2(112) - 3(70)$$

Multiply by 12

$$14(12) = 2(112 \times 12) - 3(70 \times 12)$$

$$168 = 24(112) - 36(70)$$

Thus $y_0 = 24, x_0 = -36$ is a particular solution of the given equation

The general solution is given by

$$x_1 = x_0 + \frac{b}{d} t; y_1 = y_0 - \frac{a}{d} t$$

$$x_1 = -36 + \left(\frac{112}{14}\right) t; y_1 = 24 - \left(\frac{70}{14}\right) t$$

$$x_1 = -36 + 8 t; y_1 = 24 - 5 t$$

Hence $x_1 = -36 + 8 t; y_1 = 24 - 5 t$, t is an integer

Polynomial congruence: -

$f(x) = a_0x^n + a_1x^{n-1} + a_2x^{n-2} \dots \dots \dots + a_n = 0(mod\ m)$ where $n \geq 1$ and $a_0, a_1, a_2, a_3 \dots \dots \dots a_n \in \mathbb{Z}$

Root of $f(x) \pmod m$

$x = c$ is called root of $f(x) \pmod m$ if and only if $f(c) = 0(mod\ m)$ i.e. $m \mid f(x)$

1.Find the roots of the polynomial $x^2 + 2x - 3 \pmod 5$

Consider $f(x) = x^2 + 2x - 3 \pmod 5$ $f(0) = -3 \not\equiv 0(mod\ 5)$

$f(1) = 0 \equiv 0(mod\ 5)$

$f(2) = 5 \equiv 0(mod\ 5)$

$f(3) = 12 \not\equiv 0(mod\ 5)$

$f(4) = 21 \not\equiv 0(mod\ 5)$

$x = 1, 2$ are the roots of $f(x) \pmod 5$

But $x = 0, 3, 4$ are not the roots of $f(x) \pmod 5$

SYSTEM OF LINEAR CONGRUENCE IN TWO VARIABLES

Consider a system of linear congruence in 2 variables is

$$ax + by \equiv r(\text{mod } n), \quad cx + dy \equiv s(\text{mod } n)$$

The system has a solution if $\text{GCD}(a, b, n)/c$

The system has a unique solution if $\text{GCD}(ad - bc, n) = 1$

EXAMPLES:

Solve the system of equations

$$7x + 3y \equiv 10(\text{mod } 16),$$

$$2x + 5y \equiv 9(\text{mod } 16)$$

$$a = 7, b = 3, r = 10, n = 16$$

$$c = 2, d = 5, s = 9,$$

$$\text{gcd}(a, b, n) = \text{gcd}(7, 3, 16) = 1$$

$$\text{gcd}(ad - bc, n) = 1, \quad \text{gcd}(29, 16) = 1$$

The system has a unique solution

$$ad - bc = 35 - 6 = 29$$



ATME

College of Engineering



$$7x + 3y \equiv 10(\text{mod } 16) \quad \times 5$$

$$2x + 5y \equiv 9(\text{mod } 16) \quad \times 3$$

$$29x \equiv 23(\text{mod } 16)$$

$$13x \equiv 7(\text{mod } 16)$$

$$16$$

$$\Rightarrow \frac{13x - 7}{16} \Rightarrow 13x - 7 \equiv 16k$$

$$\Rightarrow x \equiv (16k + 7)/13$$

if $k = 2$ gives integral value of x

$$x \equiv 3(\text{mod } 16)$$

$$7(3) + 3y \equiv 10(\text{mod } 16)$$

$$21 + 3y \equiv 10(\text{mod } 16)$$

$$3y \equiv -11(\text{mod } 16)$$

$$\Rightarrow 16/(3y + 11)$$

$$\Rightarrow 3y + 11 = 16k \Rightarrow y = (16k - 11)/3$$

$$\text{if } k = 2, y = 7$$

$$y \equiv 7(\text{mod } 16)$$

Wilson's Theorem

If p is a prime number, then

$$(p - 1)! \equiv -1 \pmod{p}$$

$$(p - 1)! \equiv (p - 1) \pmod{p}$$

$$(p - 2)! \equiv 1 \pmod{p}$$

$$(p - 3)! \equiv \frac{(p - 1)}{2} \pmod{p}$$

EXAMPLES

Find the remainder when $2(26)!$ is divisible by 29

$$P=29 \quad 29-26=3$$

$$(p - 3)! \equiv \frac{(p - 1)}{2} \pmod{p}$$

$$(29 - 3)! \equiv 14(\text{mod}29)$$

$$26! \equiv 14(\text{mod}29)$$

$$2(26)! \equiv 28(\text{mod}29)$$

Remainder is 28.

Euler's Theorem

For every a and n that are relatively prime $\gcd(a,n)=1$, then

$a^{\phi(n)} \equiv 1(\text{mod}n)$ Where $\phi(n)$ is Euler's totient function.

EXAMPLES

1.Solve $3^{202} \text{mod}13$ by Euler's Theorem.



$$3^{12} \equiv 1(mod13)$$

$$(3^{12})^{16} \equiv 1(mod13)$$

$$3^{192} \equiv 1(mod13) \quad (1)$$

$$3^3 \equiv 1(mod13)$$

$$(3^3)^3 \equiv 1(mod13) \quad (2)$$

$$3 \equiv 3(mod13) \quad (3)$$

By equations (1), (2), (3)

We have, $3^{202} \equiv 3(mod13)$

Fermat's Little Theorem

If p is a prime number and $(a,p)=1$ then $a^{p-1} - 1$ is divisible by p

i.e., $a^{\phi(n)} \equiv 1(mod n)$

EXAMPLES

Show that $2^{340} \equiv 1(mod 31)$ by Fermat's Little Theorem

Here $p=31$ and $a=2$

$$a^{p-1} \equiv 1(mod p)$$

$$2^{31-1} \equiv 1(mod 31)$$

$$2^{30} \equiv 1(mod 31)$$



ATME

College of Engineering



$$(2^{30})^{11} \equiv 1(\text{mod}31)$$

$$2^{330} \equiv 1(\text{mod}31)$$

$$2^5 \equiv 1(\text{mod}31)$$

$$(2^5)^2 \equiv 1^2(\text{mod}31)$$

$$2^{10} \equiv 1(\text{mod}31)$$

$$2^{330} \times 2^{10} \equiv 1(\text{mod}31)$$

$$2^{340} \equiv 1(\text{mod}31)$$



- **RSA Algorithm**
- It is an asymmetric cryptographic algorithm
- (2 keys) public and private key
- Public key: known to all users in N/W
- Private Key: Kept secret, not sharable to all
- If public key of user A is used for encryption, we have to use the private key of same user for decryption.
- The RSA scheme is a block cipher in which the plain text and cipher text are integers between 0 and $n-1$ for some n .

Key generation

Select 2 large prime numbers 'p' and 'q', $p \neq q$

Calculate $n = p * q$

Calculate $\phi(n) = (p - 1) * (q - 1)$

Choose value of e such that $1 < e < \phi(n)$ and $\gcd(\phi(n), e) = 1$

Calculate $d = e^{-1} \bmod \phi(n)$

$$de = 1 \bmod \phi(n)$$

Public key = $\{e, n\}$, Private key = $\{d, n\}$

- Encryption: $C = M^e \bmod n$

- Decryption : $M = C^d \bmod n$

EXAMPLES:

Using RSA algorithm find public key and private key with respect to $p = 3, q = 11$ and $M = 31$

- Let $p = 3, q = 11$ and $M = 31$

$$n = p * q = 3 * 11 = 33$$

$$\phi(n) = (p - 1) * (q - 1) = 2 * 10 = 20$$

So, let $e = 7$ as $1 < 7 < 20$ $\gcd(7, 20) = 1$

Now $d = e^{-1} \text{ mod } \phi(n)$

$$de = 1 \text{ mod } \phi(n)$$

$$7d = 1 \text{ mod } 20$$

$$d = 3$$



ATME

College of Engineering



Since $e = 7, d = 3$

Public key = $\{e, n\} = \{7, 33\}$

Private key = $\{d, n\} = \{3, 33\}$

Encryption :: $C = M^e \bmod n$

$$C = 31^7 \bmod n$$

$$31 = -2 \bmod 33$$

$$31^7 = -2^7 \bmod 33$$

$$31^7 = -128 \bmod 33$$

$$31^7 = -(-4) \bmod 33$$

$$31^7 = 4 \bmod 33$$

$$C = 04 = AE$$

$$M = 31 = DB$$



ATME

College of Engineering



Decryption: $M = C^d \bmod n$

$$M = 4^3 \bmod 33$$

$$4^3 = 64 \bmod 33$$

$$M = 31 \bmod 33$$

$$M = 31 = DB$$



ATME

College of Engineering



Module-2: Statistics

Curve Fitting

This is a method of finding a specific relation connecting the dependent and independent variables for a given data so as to satisfy the data as accurately as possible.

Curve fitting by the method of least squares

The method of least squares provides a relationship $y = f(x)$ such that the sum of the squares of the residuals is least. The difference between observed(y_i) and expected value (Y_i) i.e., $R_i = y_i - Y_i$ are called the residuals.

Fitting of a Straight line $y = ax + b$

Given $y = ax + b$

Normal Equations(N.E's): $a \sum x + nb = \sum y$

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

Solving the above equations we get the values of a and b , substituting these vales in the given equation leads to the curve of best fit.

Note:

Normal Equations(N.E's)for the st.line $y = a + bx$:

$$\begin{aligned}na + b \sum x &= \sum y \\a \sum x + b \sum x^2 &= \sum xy\end{aligned}$$

Problems

1. Fit a straight line $y = ax + b$ for the following data

x	1	3	4	6	8	9	11	14
y	1	2	4	4	5	7	8	9

Sol: The normal equations for fitting the Straight line $y = ax + b$ --- (*) are

$$a \sum x + nb = \sum y$$

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

Here n=8

x	y	xy	x^2
1	1	1	1
3	2	6	9
4	4	16	16
6	4	24	36
8	5	40	64
9	7	63	81
11	8	88	121
14	9	126	196
$\sum x = 56$	$\sum y = 40$	$\sum xy = 364$	$\sum x^2 = 524$

The N.E's become

$$56a + 8b = 40$$

$$524a + 56b = 364$$

On solving , we get

$$a = 0.6363 \quad \text{and} \quad b = 0.5454$$

Thus, (*) $\Rightarrow y = (0.6363)x + (0.5454)$, is the curve of best fit.

2. Find the equation of the best fitting straight line for the following data and hence estimate the value of dependent variable corresponding to the value 30 of the independent variable.

x	5	10	15	20	25
y	16	19	23	26	30

Sol: Let $y = ax + b$ --- (*) be the equation of the Straight line.
N.E's : $a \sum x + nb = \sum y$

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

Here n=5

x	y	xy	x^2
5	16	80	25
10	19	190	100
15	23	345	225
20	26	520	400
25	30	750	625
$\sum x = 75$	$\sum y = 114$	$\sum xy = 1885$	$\sum x^2 = 1375$

The N.E's become

$$75a + 5b = 114$$

$$1375a + 75b = 1885$$

On solving , we get

$$a = 0.7 \quad \text{and} \quad b = 12.3$$

Thus, (*) $\Rightarrow y = (0.7)x + (12.3)$, is the curve of best fit.

When $x = 30 \Rightarrow y = (0.7)(30) + (12.3)$

$$y = 33.3$$

3. Fit a straight line in the least square sense for the following data

x	50	70	100	120
y	12	15	21	25

Sol: Let $y = ax + b$ --- (*) be the equation of the Straight line.

N.E's : $a \sum x + nb = \sum y$

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

Here n=4

x	y	xy	x^2
50	12	600	2500
70	15	1050	4900
100	21	2100	10000
120	25	3000	14400
$\sum x = 340$	$\sum y = 73$	$\sum xy = 6750$	$\sum x^2 = 31800$

The N.E's become

$$340a + 4b = 73$$

$$13800a + 340b = 6750$$

On solving , we get

$$a = 0.19 \quad \text{and} \quad b = 2.28$$

Thus, (*) $\Rightarrow y = (0.19)x + (2.28)$, is the curve of best fit.

4. A simply supported beam carries a concentrated load P at its mid point. Corresponding to various values of P the maximum deflection Y is measured and is given in the following table.

P	100	120	140	160	180	200
Y	0.45	0.55	0.60	0.70	0.80	0.85

Find a law of the form $Y = a + bP$ and estimate Y when P is 150.

Sol: Let $Y = a + bP$ --- (*) be the equation of the Straight line.

N.E's : $na + b \sum P = \sum Y$

$$a \sum P + b \sum P^2 = \sum PY$$

Here $n = 6$

P	Y	PY	P^2
100	0.45	45	10000
120	0.55	66	14400
140	0.60	84	19600
160	0.70	112	25600
180	0.80	144	32400
200	0.85	170	40000
$\sum P = 900$	$\sum Y = 3.95$	$\sum PY = 621$	$\sum P^2 = 142000$

The N.E's become

$$6a + 900b = 3.95$$

$$900a + 142000b = 621$$

On solving , we get

$$a = 0.0476 \quad \text{and} \quad b = 0.0041$$

Thus, (*) $\Rightarrow Y = (0.0476)P + (0.0041)$, is the curve of best fit.

When $P = 150 \Rightarrow Y = (0.0476)(150) + (0.0041)$

$$Y= 0.6626$$

5. Fit a straight line to the following data.

Year	1961	1971	1981	1991	2001
Production (in tons)	8	10	12	10	16

Also find the expected production in the year 2006.

Sol: Let year and production respectively be represented by the variables ***x*** and ***y***.

Let the equation of Straight be ***y = a + bx***

Let ***X = x – 1981*** and the line of fit will be ***y = a + bX*** --- (*)

N.E's are ***na + b ∑ X = ∑ y***
a ∑ X + b ∑ X² = ∑ Xy . Here n=5

<i>x</i>	<i>X = x – 1981</i>	<i>y</i>	<i>Xy</i>	<i>X²</i>
1961	-20	8	-160	400
1971	-10	10	-100	100
1981	0	12	0	0
1991	10	10	100	100
2001	20	16	320	400
	<i>∑ X = 0</i>	<i>∑ y = 56</i>	<i>∑ Xy = 160</i>	<i>∑ X² = 1000</i>



ATME

College of Engineering



N.E's becomes

$$5a + 0.b = 56$$
$$0.a + 1000b = 160$$

On solving , we get

$$a = 11.2 \quad \text{and} \quad b = 0.16$$

Thus, (*) $\Rightarrow y = 11.2 + (0.16)X$

$$y = 11.2 + (0.16)(x - 1981)$$

$$y = -305.76 + (0.16)x, \text{ is the curve of best fit.}$$

When $x = 2006$

$$y = -305.76 + (0.16)(2006) = 15.2 \text{ tons}$$

Expected production in the year 2006 is 15.2 tons.

Fitting of a Second degree parabola $y = ax^2 + bx + c$

Given $y = ax^2 + bx + c$

Normal Equations(N.E's):

$$a \sum x^2 + b \sum x + nc = \sum y$$
$$a \sum x^3 + b \sum x^2 + c \sum x = \sum xy$$
$$a \sum x^4 + b \sum x^3 + c \sum x^2 = \sum x^2y$$

Solving the above equations we get the values of a , b and c , substituting these values in the given equation leads to the curve of best fit.

Note: If $y = a + bx + cx^2$

Normal Equations(N.E's):

$$na + b \sum x + c \sum x^2 = \sum y$$
$$a \sum x + b \sum x^2 + c \sum x^3 = \sum xy$$
$$a \sum x^2 + b \sum x^3 + c \sum x^4 = \sum x^2y$$

Problems

1. Fit a second degree parabola $y = ax^2 + bx + c$ in the least square sense for the following data and hence estimate y at $x = 6$.

x	1	2	3	4	5
y	10	12	13	16	19

Sol: Given $y = ax^2 + bx + c$

Normal Equations(N.E's):

$$\begin{aligned}a \sum x^2 + b \sum x + nc &= \sum y \\a \sum x^3 + b \sum x^2 + c \sum x &= \sum xy \\a \sum x^4 + b \sum x^3 + c \sum x^2 &= \sum x^2 y\end{aligned}$$

Here $n=5$

x	y	xy	x^2	x^2y	x^3	x^4
1	10	10	1	10	1	1
2	12	24	4	48	8	16
3	13	39	9	117	27	81
4	16	64	16	256	64	256
5	19	95	25	475	125	625
$\sum x = 15$	$\sum y = 70$	$\sum xy = 232$	$\sum x^2 = 55$	$\sum x^2y = 906$	$\sum x^3 = 225$	$\sum x^4 = 979$

The N.E's become

$$\begin{aligned}55a + 15b + 5c &= 70 \\225a + 55b + 15c &= 232 \\979a + 225b + 55c &= 906\end{aligned}$$

On Solving, we have

$$a = 0.2857 \quad b = 0.4857 \quad \text{and} \quad c = 9.4$$

Thus, (*) $\Rightarrow y = (0.2857)x^2 + (0.4857)x + (9.4)$, is the curve of best fit .

At $x = 6$,

$$y = (0.2857)(6)^2 + (0.4857)(6) + (9.4)$$

$$y = 22.78$$

2. Fit a second degree parabola $y = a + bx + cx^2$ for the data:

x	0	1	2	3	4
y	1	1.8	1.3	2.5	6.3

Sol: Given $y = a + bx + cx^2$

Normal Equations(N.E's):

$$\begin{aligned} na + b \sum x + c \sum x^2 &= \sum y \\ a \sum x + b \sum x^2 + c \sum x^3 &= \sum xy \\ a \sum x^2 + b \sum x^3 + c \sum x^4 &= \sum x^2 y \end{aligned}$$

Here $n=5$



ATME

College of Engineering



x	y	xy	x^2	x^2y	x^3	x^4
0	1	0	0	0	0	0
1	1.8	1.8	1	1.8	1	1
2	1.3	2.6	4	5.2	8	16
3	2.5	7.5	9	22.5	27	81
4	6.3	25.2	16	100.8	64	256
$\sum x = 10$	$\sum y = 12.9$	$\sum xy = 37.1$	$\sum x^2 = 30$	$\sum x^2y = 130.3$	$\sum x^3 = 100$	$\sum x^4 = 354$

The N.E's become

$$\begin{aligned}5a + 10b + 30c &= 12.9 \\10a + 30b + 100c &= 37.1 \\30a + 100b + 354c &= 130.3\end{aligned}$$

On Solving, we have

$a = 1.42$ **$b = -1.07$** and **$c = 0.55$**

Thus, (*) $\Rightarrow y = (1.42) + (-1.07)x + (0.55)x^2$, is the curve of best fit .

3. Fit a parabola $y = ax^2 + bx + c$ in the least square sense for the following data

x	2	4	6	8	10
y	3.07	12.85	31.47	57.38	91.29

Sol: Given $y = ax^2 + bx + c$
Normal Equations(N.E's):

$$\begin{aligned}
 a \sum x^2 + b \sum x + nc &= \sum y \\
 a \sum x^3 + b \sum x^2 + c \sum x &= \sum xy \\
 a \sum x^4 + b \sum x^3 + c \sum x^2 &= \sum x^2y
 \end{aligned}$$

Here n=5

x	y	xy	x^2	x^2y	x^3	x^4
2	3.07	6.14	4	12.28	8	16
4	12.85	51.40	16	205.60	64	256
6	31.47	188.82	36	1132.92	216	1296
8	57.38	459.04	64	3672.32	512	4096
10	91.29	912.90	100	9129.00	1000	10000
$\sum x = 30$	$\sum y = 196.06$	$\sum xy = 1618.3$	$\sum x^2 = 220$	$\sum x^2y = 14152.12$	$\sum x^3 = 1800$	$\sum x^4 = 15664$

The N.E's become

$$\begin{aligned}220a + 30b + 5c &= 196.06 \\1800a + 220b + 30c &= 1618.3 \\15664a + 1800b + 220c &= 14152.12\end{aligned}$$

On Solving, we have

$$a = 0.992 \qquad b = -0.8551 \quad \text{and} \qquad c = 0.696$$

Thus, (*) $\Rightarrow y = (0.992)x^2 + (-0.8551)x + (0.696)$, is the curve of best fit.

4. Find the best values of a , b , c if the equation $y = a + bx + cx^2$ is to fit most closely to the following observations

x	-2	-1	0	1	2
y	-3.150	-1.390	0.620	2.880	5.378

Sol: Equation of the parabola $y = a + bx + cx^2$

Normal Equations(N.E's):

$$na + b \sum x + c \sum x^2 = \sum y$$

$$a \sum x + b \sum x^2 + c \sum x^3 = \sum xy$$

$$a \sum x^2 + b \sum x^3 + c \sum x^4 = \sum x^2y$$

Here n=5

x	y	xy	x^2	x^2y	x^3	x^4
-2	-3.150	6.30	4	-12.60	-8	16
-1	-1.390	1.39	1	-1.39	-1	1
0	0.620	0	0	0	0	0
1	2.880	2.88	1	2.88	1	1
2	5.378	10.756	4	21.512	8	16
$\sum x = 0$	$\sum y = 4.338$	$\sum xy = 21.326$	$\sum x^2 = 10$	$\sum x^2y = 10.402$	$\sum x^3 = 0$	$\sum x^4 = 34$

The N.E's become

$$\begin{aligned}
 5a + 0.b + 10c &= 4.338 \\
 0.a + 10b + 0.c &= 21.326 \\
 10a + 0.b + 34c &= 10.402
 \end{aligned}$$

On Solving, we have

$$a = 0.6210 \qquad b = 2.1326 \qquad \text{and} \qquad c = 0.1233$$

5. Fit a parabola for the following data.

x	1	2	3	4	5	6	7	8	9
y	2	6	7	8	10	11	11	10	9

Sol: Equation of the parabola $y = a + bx + cx^2$

Normal Equations(N.E's):

$$\begin{aligned}
 na + b\sum x + c\sum x^2 &= \sum y \\
 a\sum x + b\sum x^2 + c\sum x^3 &= \sum xy \\
 a\sum x^2 + b\sum x^3 + c\sum x^4 &= \sum x^2y
 \end{aligned}$$

Here n=9



ATME

College of Engineering



x	y	xy	x^2	x^2y	x^3	x^4
1	2	2	1	2	1	1
2	6	12	4	24	8	16
3	7	21	9	63	27	81
4	8	32	16	128	64	256
5	10	50	25	250	125	625
6	11	66	36	396	216	1296
7	11	77	49	539	343	2401
8	10	80	64	640	512	4096
9	9	81	81	729	729	6561
$\sum x = 45$	$\sum y = 74$	$\sum xy = 421$	$\sum x^2 = 285$	$\sum x^2y = 2771$	$\sum x^3 = 2025$	$\sum x^4 = 15333$

The N.E's becomes

$$9a + 45b + 285c = 74$$

$$45a + 285b + 2025c = 421$$

$$285a + 2025b + 15333c = 2771$$

On Solving, we have

$$a = -0.9286 \quad b = 3.5232 \quad \text{and} \quad c = -0.2673$$

Thus, (*) $\Rightarrow y = (-0.9286) + (3.5232)x + (-0.2673)x^2$, is the curve of best fit .

6. Fit a second degree parabola to the following data

x	1	1.5	2	2.5	3	3.5	4
y	1.1	1.3	1.6	2	2.7	3.4	4.1

Sol: Equation of the parabola $y = a + bx + cx^2$

Normal Equations(N.E's):

$$\begin{aligned} na + b \sum x + c \sum x^2 &= \sum y \\ a \sum x + b \sum x^2 + c \sum x^3 &= \sum xy \\ a \sum x^2 + b \sum x^3 + c \sum x^4 &= \sum x^2y \end{aligned}$$

Here n=7

x	y	xy	x ²	x ² y	x ³	x ⁴
1	1.1	1.1	1	1.1	1	1
1.5	1.3	1.95	2.25	2.95	3.375	5.0625
2	1.6	3.2	4	6.4	8	16
2.5	2	5	6.25	12.5	15.625	39.0625
3	2.7	8.1	9	24.3	27	81
3.5	3.4	11.9	12.25	41.65	42.875	150.0625
4	4.1	16.4	16	65.6	64	256
$\sum x = 17.5$	$\sum y = 16.2$	$\sum xy = 47.65$	$\sum x^2 = 50.75$	$\sum x^2y = 154.475$	$\sum x^3 = 161.875$	$\sum x^4 = 548.1875$



The N.E's becomes

$$\begin{aligned}7a + 17.5b + 50.75c &= 16.2 \\17.5a + 50.75b + 161.875c &= 47.65 \\50.75a + 161.875b + 548.1875c &= 154.475\end{aligned}$$

On Solving, we have

$$a = 1.0357 \quad b = -0.1928 \quad \text{and} \quad c = 0.2428$$

Thus, (*) $\Rightarrow y = (1.0357) + (-0.1928)x + (0.2428)x^2$, is the curve of best fit .

Fitting of a curve of the form $y = ax^b$

Consider $y = ax^b$

Taking $\log_e$ on both sides we get,

$$\log_e y = \log_e (ax^b)$$

$$\log_e y = \log_e a + \log_e x^b$$

$$\log_e y = \log_e a + b \log_e x$$

$$Y = A + bX \quad , \text{ where } Y = \log_e y \quad , A = \log_e a \quad X = \log_e x$$

Normal Equations (N.E's) for the st. line $Y = A + bX$:

$$\begin{aligned} nA + b \sum X &= \sum Y \\ A \sum X + b \sum X^2 &= \sum XY \end{aligned}$$

Solving above equations we get A, b . But $\log_e a = A \Rightarrow a = e^A$

Problems

1. Fit a least square geometric curve $y = ax^b$ for the following data

x	1	2	3	4	5
y	0.5	2	4.5	8	12.5

Sol: Consider $y = ax^b$ --- (*)

Taking $\log_e$ on both sides we get,

$$\log_e y = \log_e(ax^b)$$

$$\log_e y = \log_e a + \log_e x^b$$

$$\log_e y = \log_e a + b \log_e x$$

$$Y = A + bX \quad , \text{ where } Y = \log_e y \quad , A = \log_e a \quad X = \log_e x$$

Normal Equations(N.E's)for the st.line $Y = A + bX$

$$\begin{aligned} nA + b \sum X &= \sum Y \\ A \sum X + b \sum X^2 &= \sum XY \end{aligned}$$

Here $n=5$

x	y	$X = \log_e x$	$Y = \log_e y$	XY	x^2
1	0.5	0	-0.6931	0	0
2	2	0.6931	0.6931	0.4804	0.4804
3	4.5	1.0986	1.5041	1.6524	1.2069
4	8	1.3863	2.0794	2.8827	1.9218
5	12.5	1.6094	2.5257	4.0649	2.5902
		$\sum X = 4.7874$	$\sum Y = 6.1092$	$\sum XY = 9.0804$	$\sum x^2 = 6.1933$

$N.E$'s becomes

$$5A + 4.7874b = 6.1092$$

$$4.7874A + 6.1993b = 9.0804$$

On Solving, we have

$$A = -0.6932 \quad b = 2$$

But $\log_e a = -0.6932 \Rightarrow a = e^{-0.6932} = 0.5$

Thus, (*) $\Rightarrow y = ax^b$

$y = (0.5)x^2$, is the curve of best fit.

2. An experiment of lifetime t of cutting tool at different cutting speeds v (units) are given below

Speed (v)	350	400	500	600
Life (t)	61	26	7	2.6

Fit a relation of the form $v = at^b$.



ATME

College of Engineering



Sol: Consider $v = at^b$ --- (*)

Taking $\log_e$ on both sides we get,

$$\log_e v = \log_e (at^b)$$

$$\log_e v = \log_e a + \log_e t^b$$

$$\log_e v = \log_e a + b \log_e t$$

$$V = A + bT, \text{ where } V = \log_e v, A = \log_e a, T = \log_e t$$

Normal Equations(N.E's)for the st.line $V = A + bT$

$$\begin{aligned} nA + b \sum T &= \sum V \\ A \sum T + b \sum T^2 &= \sum VT \end{aligned}$$

Here $n=4$

v	t	$V = \log_e v$	$T = \log_e t$	VT	T^2
350	61	5.8579	4.1109	24.0812	16.8995
400	26	5.9915	3.2581	19.5209	10.6152
500	7	6.2146	1.9459	12.093	3.7865
600	2.6	6.3969	0.9555	6.1122	0.913
		$\sum V$ $= 24.4609$	$\sum T$ $= 10.2704$	$\sum VT$ $= 61.8073$	$\sum T^2$ $= 32.2142$

N.E's becomes

$$4A + 10.2704b = 24.4609$$
$$10.2704A + 32.2142b = 61.8073$$

Solving above equations we get,

$$A = 6.5539 \quad b = -0.1709$$

But $\log_e a = 6.5539 \Rightarrow a = e^{6.5539} = 701.9766$

Thus, (*) $\Rightarrow v = at^b$

$$v = (701.9766)t^{-0.1709}, \text{ is the curve of best fit.}$$

3. Fit a curve of the form $y = ax^b$ for the following data

x	1	2	3	4	5	6
y	2.98	4.26	5.21	6.1	6.8	7.5



ATME

College of Engineering

Statistical Methods



Statistics is a science of facts and figures which may be readily available or obtained through the process of direct enquiry or enumeration. It deals with the methods of collecting, classifying and analysing the data so as to draw some valid conclusion.

‘Correlation’ can be understood as co-relation. Co-variance of two independent magnitudes is known as correlation. ‘Regression’ is an estimation of one independent variable interms of the other.

Correlation and Regression

This topic deals with data concerning independent observations.

Examples

- 1. Marks of individuals in two subjects*
- 2. Height and weight of individuals*
- 3. Amount involved in advertising a product and sale of product etc.,*

We discuss the aspect of inter-relation between the independent variables.

Correlation and Correlation Coefficient

Co-variation of two independent magnitudes is known as **correlation**. If two variables x and y are related in such a way that increase or decrease in one of them corresponds to increase or decrease in the other, we say that the variables are positively correlated. Also if increase or decrease in one of them corresponds to decrease or increase in the other, the variables are negatively correlated.

The numerical measure of correlation between two variables x and y is known as Karl Pearson's coefficient of correlation usually by r and is defined as follows

$$r = \frac{\sum XY}{\sqrt{\sum X^2} \cdot \sqrt{\sum Y^2}} \quad \text{where, } X = x - \bar{x}, Y = y - \bar{y}, \bar{x} = \frac{\sum x}{n}, \bar{y} = \frac{\sum y}{n}$$

Note:

$$1. \sigma_x^2 = \frac{\sum X^2}{n} \quad ; \quad \sigma_y^2 = \frac{\sum Y^2}{n} \quad ; \quad \sigma_x \sigma_y = \frac{\sqrt{\sum X^2} \cdot \sqrt{\sum Y^2}}{n} \quad \text{or} \quad n \sigma_x \sigma_y = \sqrt{\sum X^2} \cdot \sqrt{\sum Y^2}$$

2. The coefficient of correlation numerically does not exceed unity i.e., $-1 \leq r \leq 1$
3. If $r = \pm 1$, we say that x and y are perfectly correlated and if $r = 0$ we say that x and y are non correlated.
4.
$$r = \frac{\sigma_x^2 + \sigma_y^2 - \sigma_{x-y}^2}{2\sigma_x\sigma_y}$$

Regression

Regression is an estimation of one independent variable in terms of the other. If x and y are correlated, the best fitting straight line in the least square sense gives reasonably a good relation between x and y .

The best fitting straight line of the form $y = ax + b$ is called the regression line of y on x and $x = ay + b$ is called the regression line of x on y .

The regression line of **y on x** is

$$(y - \bar{y}) = r \frac{\sigma_y}{\sigma_x} (x - \bar{x})$$

The regression line of **x on y** is

$$(x - \bar{x}) = r \frac{\sigma_x}{\sigma_y} (y - \bar{y})$$

$r \frac{\sigma_x}{\sigma_y}$; $r \frac{\sigma_y}{\sigma_x}$ are known as regression coefficients.

Note:

1. $r = \pm \sqrt{(\text{coeff. of } x) \cdot (\text{coeff. of } y)}$, the sign of r will be positive or negative according as the regression coefficients are positive or negative.
2. The lines of regression are also of the form $Y = \frac{\sum XY}{\sum X^2} (X)$ and $X = \frac{\sum XY}{\sum Y^2} (Y)$

Problems

1. Compute the coefficient of correlation and the equation of the lines of regression for the data

x	1	2	3	4	5	6	7
y	9	8	10	12	11	13	14

Sol: Now $\bar{x} = \frac{\sum x}{n}$, $\bar{y} = \frac{\sum y}{n}$

Here $\sum x = 28$ $\sum y = 77$, $n = 7$

$$\bar{x} = \frac{28}{7}$$

$$\bar{y} = \frac{77}{7}$$

$$\bar{x} = 4$$

$$\bar{y} = 11$$

x	y	$X = x - \bar{x}$ $= x - 4$	$Y = y - \bar{y}$ $= y - 11$	XY	X^2	Y^2
1	9	-3	-2	6	9	4
2	8	-2	-3	6	4	9
3	10	-1	-1	1	1	1
4	12	0	1	0	0	1
5	11	1	0	0	1	0
6	13	2	2	4	4	4
7	14	3	3	9	9	9
				$\sum XY = 26$	$\sum X^2 = 28$	$\sum Y^2 = 28$

$$\text{Wkt, } r = \frac{\sum XY}{\sqrt{\sum X^2} \cdot \sqrt{\sum Y^2}}$$

$$r = \frac{26}{\sqrt{28} \cdot \sqrt{28}}$$

$$r = \frac{26}{28}$$
$$r = \mathbf{0.9286}$$

The lines regression are given by

$$(y - \bar{y}) = r \frac{\sigma_y}{\sigma_x} (x - \bar{x})$$

$$(y - 11) = 0.9286(x - 4)$$

$$r \frac{\sigma_y}{\sigma_x} = \frac{\sum XY}{\sum X^2} = \frac{26}{28} = \mathbf{0.9286}$$

$$y - 11 = 0.9286x - 3.7144$$

$$y = 0.9286x + 7.2856$$

$$\text{Also, } (x - \bar{x}) = r \frac{\sigma_x}{\sigma_y} (y - \bar{y})$$

$$(x - 4) = 0.9286(y - 11)$$

$$x - 4 = 0.9286y - 10.2146$$

$$x = 0.9286y - 6.2146$$

$$r \frac{\sigma_x}{\sigma_y} = \frac{\sum XY}{\sum Y^2} = \frac{26}{28} = 0.9286$$

2. Obtain the lines of regression and hence find the coefficient of correlation for the following data

x	1	2	3	4	5	6	7
y	9	8	10	12	11	13	14

Sol: Now $\bar{x} = \frac{\sum x}{n}$, $\bar{y} = \frac{\sum y}{n}$

Here $\sum x = 28$ $\sum y = 77$, $n = 7$, $\bar{x} = \frac{28}{7}$, $\bar{x} = 4$; $\bar{y} = \frac{77}{7}$, $\bar{y} = 11$

x	y	$X = x - \bar{x}$ $= x - 4$	$Y = y - \bar{y}$ $= y - 11$	XY	X^2	Y^2
1	9	-3	-2	6	9	4
2	8	-2	-3	6	4	9
3	10	-1	-1	1	1	1
4	12	0	1	0	0	1
5	11	1	0	0	1	0
6	13	2	2	4	4	4
7	14	3	3	9	9	9
				$\sum XY = 26$	$\sum X^2 = 28$	$\sum Y^2 = 28$

The lines regression are given by

$$(y - \bar{y}) = \frac{\sum XY}{\sum X^2} (x - \bar{x})$$

$$(y - 11) = 0.9286(x - 4)$$

$$y - 11 = 0.9286x - 3.7144$$

$$y = \mathbf{0.9286x + 7.2856}$$

$$\frac{\sum XY}{\sum X^2} = \frac{26}{28} = \mathbf{0.9286}$$

Also , $(x - \bar{x}) = \frac{\sum XY}{\sum Y^2} (y - \bar{y})$

$$(x - 4) = 0.9286(y - 11)$$

$$x - 4 = 0.9286y - 10.2146$$

$$x = \mathbf{0.9286y - 6.2146}$$

$$\frac{\sum XY}{\sum Y^2} = \frac{26}{28} = \mathbf{0.9286}$$

Wkt $r = \sqrt{(\text{coeff. of } x) \cdot (\text{coeff. of } y)}$

r must be positive since regression coefficients are +ve

$$r = \sqrt{(0.9286) \cdot (0.9286)}$$

$$r = 0.9286$$

3. Find the correlation coefficient for the following two groups

A	92	89	87	86	83	77	71	63	53	50
B	86	83	91	77	68	85	52	82	37	57

Sol: Now $\bar{x} = \frac{\sum x}{n}$, $\bar{y} = \frac{\sum y}{n}$

Here $\sum x = 751$ $\sum y = 718$, $n = 10$

$$\bar{x} = \frac{751}{10} , \bar{x} = 75.1 \quad ; \quad \bar{y} = \frac{718}{10} , \bar{y} = 71.8$$

x	y	$X = x - \bar{x}$ $= x - 75.1$	$Y = y - \bar{y}$ $= y - 71.8$	XY	X^2	Y^2
92	86	16.9	14.2	239.98	285.61	201.64
89	83	13.9	11.2	155.68	193.21	125.44
87	91	11.9	19.2	228.48	141.61	368.64
86	77	10.9	5.2	56.68	118.81	27.04
83	68	7.9	-3.8	-30.02	62.41	14.44
77	85	1.9	13.2	25.08	3.61	174.24
71	52	-4.1	-19.8	81.18	16.81	392.04
63	82	-12.1	10.2	-123.42	146.41	104.04
53	37	-22.1	-34.8	769.08	488.41	1211.04
50	57	-25.1	-14.8	371.48	630.01	219.04
				$\sum XY = 1774.2$	$\sum X^2 = 2086.59$	$\sum Y^2 = 2837.6$

$$\text{Wkt, } r = \frac{\sum XY}{\sqrt{\sum X^2} \cdot \sqrt{\sum Y^2}}$$

$$r = \frac{1774.2}{\sqrt{2086.59} \cdot \sqrt{2837.6}}$$

$$r = \frac{1774.2}{2433.29}$$

$$r = 0.7291$$

4. Find the correlation coefficient and the equation of the lines of regression for the following values of x and y

x	1	2	3	4	5
y	2	5	3	8	7

Sol: Now $\bar{x} = \frac{\sum x}{n}$, $\bar{y} = \frac{\sum y}{n}$

Here $\sum x = 15$ $\sum y = 25$, $n = 5$

$\bar{x} = \frac{15}{5}$, $\bar{x} = 3$; $\bar{y} = \frac{25}{5}$, $\bar{y} = 5$

x	y	$X = x - \bar{x}$ $= x - 3$	$Y = y - \bar{y}$ $= y - 5$	XY	X^2	Y^2
1	2	-2	-3	6	4	9
2	5	-1	0	0	1	0
3	3	0	-2	0	0	4
4	8	1	3	3	1	9
5	7	2	2	4	4	4
				$\sum XY = 13$	$\sum X^2 = 10$	$\sum Y^2 = 26$

$$\text{Wkt, } r = \frac{\sum XY}{\sqrt{\sum X^2} \cdot \sqrt{\sum Y^2}}$$

$$r = \frac{13}{\sqrt{10} \cdot \sqrt{26}}$$

$$r = 0.8062$$

$$\text{Now, } \sigma_x^2 = \frac{\sum X^2}{n} = \frac{10}{5} = 2 \quad ; \quad \sigma_x^2 = 2 \quad ; \quad \sigma_x = \sqrt{2}$$

$$\sigma_y^2 = \frac{\sum Y^2}{n} = \frac{26}{5} = 5.2 \quad ; \quad \sigma_y^2 = 5.2 \quad ; \quad \sigma_y = \sqrt{5.2}$$

The lines regression are given by

$$(y - \bar{y}) = r \frac{\sigma_y}{\sigma_x} (x - \bar{x})$$



ATME

College of Engineering



$$(y - 5) = (0.8062) \frac{\sqrt{2}}{\sqrt{5.2}} (x - 3)$$

$$(y - 5) = (1.3) (x - 3)$$

$$y - 5 = 1.3x - 3.9$$

$$y = 1.3x - 3.9 + 5$$

$$\mathbf{y = 1.3x + 1.1}$$

$$\text{Also, } (x - \bar{x}) = r \frac{\sigma_x}{\sigma_y} (y - \bar{y})$$

$$(x - 3) = (0.8062) \frac{\sqrt{2}}{\sqrt{5.2}} (y - 5)$$

$$x - 3 = 0.5(y - 5)$$

$$x = 0.5y - 2.5 + 3$$

$$\mathbf{x = 0.5y + 0.5}$$

5. The following data gives the age of husband (x) and the age of wife (y) in years. Form the two regression lines and calculate the age of husband corresponding to 16 years age of wife.

x	36	23	27	28	28	29	30	31	33	35
y	29	18	20	22	27	21	29	27	29	28

Sol: Now $\bar{x} = \frac{\sum x}{n}$, $\bar{y} = \frac{\sum y}{n}$

Here $\sum x = 300$ $\sum y = 250$, $n = 10$

$$\bar{x} = \frac{300}{10}$$

$$\bar{y} = \frac{250}{10}$$

$$\bar{x} = 30$$

$$\bar{y} = 25$$

x	y	$X = x - \bar{x}$ $= x - 30$	$Y = y - \bar{y}$ $= y - 25$	XY	X^2	Y^2
36	29	6	4	24	36	16
23	18	-7	-7	49	49	49
27	20	-3	-5	15	9	25
28	22	-2	-3	6	4	9
28	27	-2	2	-4	4	4
29	21	-1	-4	4	1	16
30	29	0	4	0	0	16
31	27	1	2	2	1	4
33	29	3	4	12	9	16
35	28	5	3	15	25	9
				$\sum XY = 123$	$\sum X^2 = 138$	$\sum Y^2 = 164$

The lines regression are given by

$$(y - \bar{y}) = \frac{\sum XY}{\sum X^2} (x - \bar{x})$$

$$(y - 25) = 0.8913(x - 30)$$

$$y - 25 = 0.8913x - 26.739$$

$$y = 0.8913x - 1.739$$

Also , $(x - \bar{x}) = \frac{\sum XY}{\sum Y^2} (y - \bar{y})$

$$(x - 30) = 0.75(y - 25)$$

$$x - 30 = 0.75y - 18.75$$

$$x = 0.75y + 11.25$$

$$\frac{\sum XY}{\sum X^2} = \frac{123}{138} = 0.8913$$

$$\frac{\sum XY}{\sum Y^2} = \frac{123}{164} = 0.75$$

When $y = 16$, $x = 0.75(16) + 11.25$ $x = 23.25 \cong 23$

Thus husband's age is 23 years corresponding to wife age of 16 years.

6. In a partially destroyed lab record, only the lines of regression of y on x and x on y are available as $4x - 5y + 33 = 0$ and $20x - 9y = 107$ respectively. Calculate the coefficient of correlation between x and y .

Sol: Lines of regression of y on x and x on y are

$$4x - 5y + 33 = 0$$

$$20x - 9y = 107$$

$$5y = 4x + 33$$

$$20x = 9y + 107$$

$$y = \frac{4}{5}x + \frac{33}{5}$$

$$x = \frac{9}{20}y + \frac{107}{20}$$

$$\text{Coefficient of } x = \frac{4}{5}$$

$$\text{Coefficient of } y = \frac{9}{20}$$

The coefficient of correlation is

$$r = \sqrt{(\text{coeff. of } x) \cdot (\text{coeff. of } y)}$$

$$r = \sqrt{\left(\frac{4}{5}\right) \cdot \left(\frac{9}{20}\right)}$$

r must be positive since regression coefficients are +ve



ATME

College of Engineering



$$r = \sqrt{\frac{36}{100}}$$

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

$$\mathbf{r = 0.6}$$

7. Given the equation of the regression lines $x = 19.13 - 0.87y$, $y = 11.64 - 0.5x$.

Compute the mean of x 's , mean of y 's and the coefficient of correlation.

Sol: To find $\bar{x}$, $\bar{y}$ and r .

We know that the regression lines passes through $(\bar{x}, \bar{y})$.

$$\therefore \bar{x} = 19.13 - 0.87\bar{y} \quad \text{and} \quad \bar{y} = 11.64 - 0.5\bar{x}$$

That is , $\bar{x} + 0.87\bar{y} = 19.13$
 $0.5\bar{x} + \bar{y} = 11.64$

Solving above equations , we get

$\bar{x} = 15.94$ and $\bar{y} = 3.67$

Rewriting the equations

$y = -0.5x + 11.64$;

Coefficient of $x = -0.5$

$x = -0.87y + 19.13$

Coefficient of $y = -0.87$

The coefficient of correlation is

$r = -\sqrt{(\text{coeff. of } x) \cdot (\text{coeff. of } y)}$ r must be negative since regression coefficients are –ve

$r = -\sqrt{(-0.5) \cdot (-0.87)}$

$r = -0.6595$

8. Compute $\bar{x}$, $\bar{y}$ and r from the following equation of the regression lines $2x + 3y + 1 = 0$ and $x + 6y - 4 = 0$

Sol: We know that the regression lines passes through $(\bar{x}, \bar{y})$.

$$\therefore 2\bar{x} + 3\bar{y} = -1 \quad \text{and} \quad \bar{x} + 6\bar{y} = 4$$

Solving above equations , we get

$$\bar{x} = -2 \quad \text{and} \quad \bar{y} = 1$$

Consider the regression lines

$$2x + 3y + 1 = 0 \quad \text{and}$$

$$x + 6y - 4 = 0$$

$$3y = -2x - 1 \quad \text{and}$$

$$x = -6y + 4$$

$$y = -\frac{2}{3}x - \frac{1}{3}$$

$$\text{Coefficient of } x = -\frac{2}{3}$$

$$\text{Coefficient of } y = -6$$

The coefficient of correlation is

$$r = -\sqrt{(\text{coeff. of } x) \cdot (\text{coeff. of } y)} \quad r \text{ must be negative since regression coefficients are } -ve$$

$$r = -\sqrt{\left(-\frac{2}{3}\right) \cdot (-6)}$$

$$r = -\sqrt{4}$$

$r = -2$, this is absurd. Since r lies between -1 and +1.

Rewriting the regression lines

$$x + 6y - 4 = 0 \quad \text{and}$$

$$2x + 3y + 1 = 0$$

$$6y = -x + 4 \quad \text{and}$$

$$2x = -3y - 1$$

$$y = -\frac{1}{6}x + \frac{4}{6} \quad \text{and}$$

$$x = -\frac{3}{2}y - \frac{1}{2}$$

Coefficient of $x = -\frac{1}{6}$

Coefficient of $y = -\frac{3}{2}$

The coefficient of correlation is

$$r = -\sqrt{(\text{coeff. of } x) \cdot (\text{coeff. of } y)}$$

r must be negative since regression coefficients are –ve

$$r = -\sqrt{\left(-\frac{1}{6}\right) \cdot \left(-\frac{3}{2}\right)}$$

$$r = -\sqrt{\frac{1}{4}}$$

$$r = -0.5$$

Thus, $\bar{x} = -2$, $\bar{y} = 1$ and $r = -0.5$

9. Given _____ and $r = 0.8$, write down the equation of the lines of

	x-series	y-series
Mean	18	100
SD	14	20

regression and hence find the most probable value of y when $x = 70$.

Sol: By data $\bar{x} = 18$ and $\bar{y} = 100$; $\sigma_x = 14$ and $\sigma_y = 20$; $r = 0.8$

The equations of regression lines are

$$(y - \bar{y}) = r \frac{\sigma_y}{\sigma_x} (x - \bar{x})$$

$$(y - 100) = (0.8) \frac{20}{14} (x - 18)$$

$$y - 100 = 1.14(x - 18)$$

$$y = 1.14x + 79.48$$

$$(x - \bar{x}) = r \frac{\sigma_x}{\sigma_y} (y - \bar{y})$$

$$(x - 18) = (0.8) \frac{14}{20} (y - 100)$$

$$x - 18 = 0.56(y - 100)$$

$$x = 0.56y - 38$$

When $x = 70$, $y = 1.14(70) + 79.48$; $y = 159.28$

10. Show that if θ is the angle between the lines of regression, then $\tan\theta = \frac{\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2} \left(\frac{1-r^2}{r} \right)$

Sol:

We know that if θ is acute, the angle between the lines $y = m_1x + c_1$ and $y = m_2x + c_2$ is given by,

$$\tan\theta = \frac{m_2 - m_1}{1 + m_1m_2} \quad \text{--- (*)}$$

The equations of regression lines are ,

$$(y - \bar{y}) = r \frac{\sigma_y}{\sigma_x} (x - \bar{x}) \quad \text{--- (1)} \quad \text{and} \quad (x - \bar{x}) = r \frac{\sigma_x}{\sigma_y} (y - \bar{y}) \quad \text{--- (2)}$$

We rewrite the equation (2) as

$$(y - \bar{y}) = \frac{\sigma_y}{r \cdot \sigma_x} (x - \bar{x})$$

Slopes of (1) and (2) are $m_1 = r \frac{\sigma_y}{\sigma_x}$ and $m_2 = \frac{\sigma_y}{r \cdot \sigma_x}$



ATME

College of Engineering



Substituting the slopes in (*), we get

$$\tan\theta = \frac{\frac{\sigma_y}{r \cdot \sigma_x} - r \frac{\sigma_y}{\sigma_x}}{1 + r \frac{\sigma_y}{\sigma_x} \cdot \frac{\sigma_y}{r \cdot \sigma_x}}$$

$$\tan\theta = \frac{\frac{\sigma_y}{\sigma_x} \left(\frac{1}{r} - r \right)}{1 + r \cdot \frac{1}{r} \frac{\sigma_y^2}{\sigma_x^2}}$$

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

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

$$\tan\theta = \frac{\sigma_y}{\sigma_x} \cdot \frac{\sigma_x^2}{(\sigma_x^2 + \sigma_y^2)} \left(\frac{1 - r^2}{r} \right)$$

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

Note: If $r = \pm 1$, $\tan\theta = 0 \Rightarrow \theta = 0$, which implies that the two regression lines coincide and hence the variables are perfectly correlated. Also if $r = 0$, $\tan\theta = \infty \Rightarrow \theta = \frac{\pi}{2}$, which implies that the lines are perpendicular and hence the variables are uncorrelated.

11. In a bivariate distribution $\sigma_x = \sigma_y$ and the angle between the regression lines is $\tan^{-1}(3)$. Find the correlation coefficient.

Sol: Let θ be the angle between the lines of regression, we have

$$\tan\theta = \frac{\sigma_x\sigma_y}{(\sigma_x^2 + \sigma_y^2)} \left(\frac{1-r^2}{r} \right) \quad \text{---- (1)}$$

By data, $\theta = \tan^{-1}(3)$ or $\tan\theta = 3$ and $\sigma_x = \sigma_y$

$$\text{Hence (1)} \Rightarrow 3 = \frac{\sigma_x \cdot \sigma_x}{(\sigma_x^2 + \sigma_x^2)} \left(\frac{1-r^2}{r} \right)$$

$$3 = \frac{\sigma_x^2}{2\sigma_x^2} \left(\frac{1-r^2}{r} \right)$$

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

$$6r = 1 - r^2$$



ATME

College of Engineering



$$r^2 + 6r - 1 = 0$$

$$r = \frac{-6 \pm \sqrt{6^2 - 4 \cdot 1 \cdot (-1)}}{2 \cdot 1}$$

$$r = \frac{-6 \pm \sqrt{36 + 4}}{2}$$

$$r = \frac{-6 \pm \sqrt{40}}{2}$$

$$r = \frac{-6 \pm 2\sqrt{10}}{2}$$

$$r = -3 \pm \sqrt{10}$$

$$r = +0.1623 \quad \text{or} \quad r = -6.1623(\text{neglect})$$

Since $-1 \leq r \leq 1$, **$r = 0.1623$**

12. If the coefficient of correlation between two variables x and y is 0.5 and the acute angle between their lines of regressions is $\tan^{-1} \left(\frac{3}{5} \right)$, show that $\sigma_x = 2\sigma_y$ or $\sigma_y = 2\sigma_x$.

Sol: By data $r = 0.5$ and $\theta = \tan^{-1} \left(\frac{3}{5} \right)$ or $\tan \theta = \frac{3}{5}$

The angle between the lines of regression is given by

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

$$\frac{3}{5} = \frac{\sigma_x \sigma_y}{(\sigma_x^2 + \sigma_y^2)} \left(\frac{1 - \left(\frac{1}{2} \right)^2}{\frac{1}{2}} \right)$$



ATME

College of Engineering



$$\frac{3}{5} = \frac{\sigma_x \sigma_y}{(\sigma_x^2 + \sigma_y^2)} \left(\frac{1 - \frac{1}{4}}{\frac{1}{2}} \right)$$

$$\frac{3}{5} = \frac{\sigma_x \sigma_y}{(\sigma_x^2 + \sigma_y^2)} \left(\frac{3}{4} \cdot \frac{2}{1} \right)$$

$$\frac{3}{5} = \frac{3\sigma_x \sigma_y}{2(\sigma_x^2 + \sigma_y^2)}$$

$$\frac{1}{5} = \frac{\sigma_x \sigma_y}{2(\sigma_x^2 + \sigma_y^2)}$$

$$2(\sigma_x^2 + \sigma_y^2) = 5\sigma_x \sigma_y$$



ATME

College of Engineering



$$2\sigma_x^2 + 2\sigma_y^2 = 5\sigma_x\sigma_y$$

$$2\sigma_x^2 - 5\sigma_x\sigma_y + 2\sigma_y^2 = 0$$

$$2\sigma_x^2 - \sigma_x\sigma_y - 4\sigma_x\sigma_y + 2\sigma_y^2 = 0$$

$$\sigma_x(2\sigma_x - \sigma_y) - 2\sigma_y(2\sigma_x - \sigma_y) = 0$$

$$(2\sigma_x - \sigma_y)(\sigma_x - 2\sigma_y) = 0$$

Thus, we get

$$\sigma_y = 2\sigma_x \quad \text{or} \quad \sigma_x = 2\sigma_y$$



ATME

College of Engineering



Module-3: Probability Distributions



Introduction

The three main components of statistics are collection of data, analysis of data and inference. The scope of statistics is very vast and finds its applications in almost all spheres. Statistical methods guides us in any scientific enquiry and leans heavily on probability theory.

Basic Definitions and concepts

Experiment is a process by which measurement or observation is obtained.

Random experiment : An experiment which does not result in the same out come when performed under the same conditions.

Eg : 1. Tossing of a fair coin.

2. Throwing of a fair die.



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

Eg 1. $S = \{H, T\}$ is the sample space in tossing of a single coin.

2. $S = \{HH, TH, HT, TT\}$ is the sample space obtained in tossing of two coins.

3. $S = \{1, 2, 3, 4, 5, 6\}$ in throwing of a die.

Event : Any event is a subset of an appropriate sample space.

Exhaustive event : An event consisting of all the various possibilities is called an exhaustive event.

Mutually exclusive events : Events are said to be mutually exclusive if and only if one of them can take place at a time, in other words happening of one event prevent the simultaneous happening of the others.

Eg: In the tossing of a coin, either heads or tails may turn up but not both.



Independent events: Two or more events are said to be independent if the happening or non happening of one event does not prevent the happening or non happening of the others.

Eg: 1. When two coins are tossed the event of getting head is an independent event as both the coins can turn out heads.

Probability is the measure of the chance with which we can expect the event to occur. If an event A can occur in **m** different ways out of number of **n** possible ways, all of which are equally likely then the probability of the event A denoted by P(A) is given by $P(A) = \frac{m}{n}$.

Note: 1) Since $0 \leq m \leq n \Rightarrow P(A)$ is non negative and $0 \leq P(A) \leq 1$

2) If $P(A) = 0$ then A is called a null event.

3) If $P(A) = 1$ then A is called a sure event.

4) $P(A) + P(\bar{A}) = 1$, the set of all unfavourable events is denoted by $\bar{A}$.



Addition theorem of probability

The probability of the happening of one or the other mutually exclusive events is equal to the sum of the probabilities of the two events.

That is, if A, B are two mutually exclusive events then,

$$P(A \text{ or } B) = P(A) + P(B)$$

Product theorem of probability

If a compound event is made up of a number of independent events, the probability of the happening of the compound event is equal to the product of the probabilities of the independent events .

That is, if A, B are independent events then,

$$P(A \text{ and } B) = P(A).P(B)$$



RANDOM VARIABLE & PROBABILITY DISTRIBUTION

Introduction

When we consider an experiment of tossing two coins. The sample points or outcomes are qualitative and are described by their attributes though the probabilities attached to each of these outcomes are quantitative. We now associate to each sample point a real number called the **random variable**

Random variable

Definition: Let S be a sample space then $X : S \rightarrow \mathbf{R}$ defined over the elements of S is called a **random variable**. A random variable is a function that assigns a real number to every sample point in the sample space of a random experiment. Random variables are usually denoted by $X, Y, Z \dots$ and it may be noted that different random variables may be associated with the sample space S . The set of all real numbers of a random variable X is called the range of X .

Eg:

1. While tossing a coin, suppose that the value 1 is associated for the outcome 'head' and 0 for the outcome tail. We have the sample space $S = \{H, T\}$ and if X is the random variable, then $X(H) = 1$ & $X(T) = 0$.

Range of $X = \{0,1\}$

2. Suppose a coin is tossed twice, we shall associate two different random variables X, Y as follows, where we have the sample space

$S = \{HH, HT, TH, TT\}$

X = Number of 'heads' in the outcome.

The association of the elements in S to X is as follows.

Out come	HH	HT	TH	TT
Random variable X	2	1	1	0

Range of $X = \{0,1,2\}$

Suppose $Y =$ Number of 'tails' in the outcome, we have

Out come	HH	HT	TH	TT
Random variable Y	0	1	1	2

Range of $Y = \{0,1,2\}$

Discrete and Continuous random variables

If a random variable takes finite or countably infinite number of values then it is called a *discrete random variable*.

Discrete random variable will have finite or countably infinite range.

Eg: 1. Tossing of a coin and observing the outcome.

2. Throwing a 'die' and observing the numbers on the face.

If a random variable takes non countable number of values then it is called a *non discrete or continuous random variable*.

Eg:

1. Weight of articles
2. Length of nails produced by machine.

Generally counting problems correspond to discrete random variables and measuring problems lead to continuous random variables.

Probability function and Discrete probability distribution

If for each value x_i of a discrete random variable X , we assign a real number $p(x_i)$ such that

i) $p(x_i) \geq 0$ ii) $\sum_i p(x_i) = 1$ then the function $p(x)$ is called a *probability function*.

If the probability that X take the values x_i is p_i , then $P(X = x_i) = p_i$ or $p(x_i)$.

The set of values $[x_i, p(x_i)]$ is called a discrete (finite) probability distribution of the discrete random variable X . The function $P(X)$ is called the *probability density function (p.d.f)* or the *probability mass function (p.m.f)*

The distribution function $f(x)$ defined by $f(x) = P(X \leq x) = \sum_{i=1}^n P(x_i)$, x being an integer, is called the *cumulative distribution function (c.d.f)*.

The mean and variance of the discrete probability distribution are defined as follows

$$\text{Mean } \mu = \sum_i x_i \cdot p(x_i)$$

$$\text{Variance } V = \sum_i (x_i - \mu)^2 \cdot p(x_i)$$

$$V = \sum_i x_i^2 \cdot p(x_i) - [\sum_i x_i \cdot p(x_i)]^2$$

$$V = \sum x_i^2 \cdot p(x_i) - \mu^2$$

$$\text{Standard deviation } \sigma = \sqrt{V}$$

Problems

1. A coin is tossed twice. A random variable X represent the number of heads turning up. Find the discrete probability distribution for X . Also find its mean and variance.

Sol: Here $S = \{HH, HT, TH, TT\}$
 $X = \{0, 1, 2\}$

$$\text{Now, } P(HH) = \frac{1}{4} \quad P(HT) = \frac{1}{4} \quad P(TH) = \frac{1}{4} \quad P(TT) = \frac{1}{4}$$

$$P(X = 0, \text{ i.e., no head turning up}) = P(TT) = \frac{1}{4}$$

$$P(X = 1, \text{ i.e., one head turning up}) = P(HT \cup TH) = P(HT) + P(TH) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$$

$$P(X = 2, \text{ i.e., two heads turning up}) = P(HH) = \frac{1}{4}$$

The discrete probability distribution for X is as follows

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

We observe that, $p(x_i) > 0$ and $\sum_i p(x_i) = 1$.

$$\text{Mean, } \mu = \sum_i x_i \cdot p(x_i) = (0) \cdot \left(\frac{1}{4}\right) + (1) \cdot \left(\frac{1}{2}\right) + (2) \cdot \left(\frac{1}{4}\right) = \frac{1}{2} + \frac{1}{2}$$

$$\mu = 1$$

$$\text{Variance, } V = \sum_i (x_i - \mu)^2 \cdot p(x_i)$$

$$V = (0 - 1)^2 \cdot \frac{1}{4} + (1 - 1)^2 \cdot \frac{1}{2} + (2 - 1)^2 \cdot \frac{1}{4}$$

$$V = \frac{1}{4} + 0 + \frac{1}{4}$$

$$V = \frac{1}{2}$$

2. A random experiment of tossing a 'die' twice is performed. Random variable X on this sample space is defined to be the sum of the two numbers turning up on the toss. Find the discrete probability distribution for the random variable X and compute the corresponding mean and standard deviation.

Sol: Here $S = \{(x, y) \text{ where } x = 1, 2, \dots, 6 ; y = 1, 2, \dots, 6\}$

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

Number of elements in $S=36$.

The set of values of the random variable X defined as sum of two numbers on the face of the 'die' are $\{2, 3, 4, 5, \dots, 10, 11, 12\}$

Now, $p(x_1) = \frac{n(E)}{n(S)} = \frac{1}{36}$; $p(x_2) = \frac{2}{36}$; ...

The discrete probability distribution for X is as follows

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

Here $p(x_i) > 0$ and $\sum_i p(x_i) = 1$.

Mean $\mu = \sum_i x_i \cdot p(x_i)$

$$\mu = (2) \cdot \left(\frac{1}{36}\right) + (3) \cdot \left(\frac{2}{36}\right) + \dots + (11) \cdot \left(\frac{2}{36}\right) + (12) \cdot \left(\frac{1}{36}\right)$$

$$\mu = \frac{252}{36} = 7$$

Variance, $V = \sum_i (x_i - \mu)^2 \cdot p(x_i)$

$$V = (2 - 7)^2 \cdot \frac{1}{36} + (3 - 7)^2 \cdot \frac{2}{36} + \dots + (12 - 7)^2 \cdot \frac{1}{36}$$

$$V = \frac{210}{36} = \frac{35}{6}$$

Std. deviation, $\sigma = \sqrt{V}$

$$\sigma = \sqrt{\frac{35}{6}} = 2.42$$

3. Show that the following distribution represents a discrete probability distribution. Find the mean and variance

x	10	20	30	40
$p(x)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

Sol: We observe that $p(x) > 0$ and $\sum p(x) = 1$.

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

$$\mu = (10) \cdot \frac{1}{8} + (20) \cdot \frac{3}{8} + (30) \cdot \frac{3}{8} + (40) \cdot \frac{1}{8}$$

$$\mu = \frac{10+60+90+40}{8} = \frac{200}{8}$$

$$\mu = 25$$

Variance, $V = \sum (x - \mu)^2 \cdot p(x)$

$$V = (10 - 25)^2 \cdot \frac{1}{8} + (20 - 25)^2 \cdot \frac{3}{8} + (30 - 25)^2 \cdot \frac{3}{8} + (40 - 25)^2 \cdot \frac{1}{8}$$

$$V = \frac{600}{8} = 75$$

4. Find the value of k such that the following distribution represents a finite probability distribution. Hence find its mean and standard deviation. Also find $p(x \leq 1)$, $p(x > 1)$ and $p(-1 < x \leq 2)$

x	-3	-2	-1	0	1	2	3
$p(x)$	k	$2k$	$3k$	$4k$	$3k$	$2k$	k

Sol: We must have $p(x) \geq 0$ for all x and $\sum p(x) = 1$. The first condition is satisfied if $k \geq 0$
Since $\sum p(x) = 1$

$$k + 2k + 3k + 4k + 3k + 2k + k = 1$$

$$16k = 1$$

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

The discrete probability distribution is as follows

x	-3	-2	-1	0	1	2	3
$p(x)$	$\frac{1}{16}$	$\frac{2}{16}$	$\frac{3}{16}$	$\frac{4}{16}$	$\frac{3}{16}$	$\frac{2}{16}$	$\frac{1}{16}$

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

$$\mu = (-3) \cdot \frac{1}{16} + (-2) \cdot \frac{2}{16} + (-1) \cdot \frac{3}{16} + (0) \cdot \frac{4}{16} + (1) \cdot \frac{3}{16} + (2) \cdot \frac{2}{16} + (3) \cdot \frac{1}{16}$$

$$\mu = 0$$

Variance, $V = \sum (x - \mu)^2 \cdot p(x)$

$$V = (-3 - 0)^2 \cdot \frac{1}{16} + (-2 - 0)^2 \cdot \frac{2}{16} + (-1 - 0)^2 \cdot \frac{3}{16} + (0 - 0)^2 \cdot \frac{4}{16} + (1 - 0)^2 \cdot \frac{3}{16} + (2 - 0)^2 \cdot \frac{2}{16} + (3 - 0)^2 \cdot \frac{1}{16}$$

$$V = \frac{1}{16} (9 + 8 + 3 + 0 + 3 + 8 + 9) = \frac{40}{16} = \frac{5}{2}$$

Standard deviation $\sigma = \sqrt{V}$

$$\sigma = \sqrt{\frac{5}{2}} = 1.58$$

Now, $p(x \leq 1) = p(-3) + p(-2) + p(-1) + p(0) + p(1)$

$$= \frac{1}{16} + \frac{2}{16} + \frac{3}{16} + \frac{4}{16} + \frac{3}{16}$$

$$p(x \leq 1) = \frac{13}{16}$$

$p(x > 1) = p(2) + p(3)$

$$= \frac{2}{16} + \frac{1}{16}$$

$$p(x > 1) = \frac{3}{16}$$

$p(-1 < x \leq 2) = p(0) + p(1) + p(2)$

$$= \frac{4}{16} + \frac{3}{16} + \frac{2}{16}$$

$$p(-1 < x \leq 2) = \frac{9}{16}$$

5. A random variable X has the following probability function for various values of x .

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

i) Find k ii) Evaluate $P(x < 6)$, $P(x \geq 6)$ and $P(3 < x \leq 6)$. Also find the probability distribution and the distribution function of X .

Sol: We must have $P(x) \geq 0$ and $\sum P(x) = 1$.

The first condition is satisfied for $k \geq 0$ and we have to find k such that

$$\sum P(x) = 1$$

$$\text{i.e, } 0 + k + 2k + 2k + 3k + k^2 + 2k^2 + (7k^2 + k) = 1$$

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

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

$$k = \frac{1}{10} \quad \text{and} \quad k = -1 \text{ (Neglect)}$$

Hence we have the following table

x	0	1	2	3	4	5	6	7
$P(x)$	0	$\frac{1}{10}$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{3}{10}$	$\frac{1}{100}$	$\frac{1}{50}$	$\frac{17}{100}$

Now, $P(x < 6) = P(0) + P(1) + P(2) + P(3) + P(4) + P(5)$

$$= 0 + \frac{1}{10} + \frac{1}{5} + \frac{1}{5} + \frac{3}{10} + \frac{1}{100} = \frac{81}{100}$$

$$P(x < 6) = \mathbf{0.81}$$

$$P(x \geq 6) = P(6) + P(7)$$

$$P(x \geq 6) = \frac{1}{50} + \frac{17}{100} = \frac{19}{100}$$

$$P(x \geq 6) = \mathbf{0.19}$$

$$\begin{aligned}
 P(3 < x \leq 6) &= P(4) + P(5) + P(6) \\
 &= \frac{3}{10} + \frac{1}{100} + \frac{1}{50} = \frac{33}{100}
 \end{aligned}$$

$P(3 < x \leq 6) = 0.33$

The probability distribution is as follows

x	0	1	2	3	4	5	6	7
$P(x)$	0	0.1	0.2	0.2	0.3	0.01	0.02	0.17

The distribution function of Y is $f(x) = P(Y \leq x) = \sum^n_{i=1} p(x_i)$ is called cumulative

x	0	1	2	3	4	5	6	7
$f(x)$	0	0+0.1 =0.1	0.1+0.2 =0.3	0.3+0.2 =0.5	0.5+0.3 =0.8	0.8+0.01 =0.81	0.81+0.02 =0.83	0.83+0.17 =1

6. A random variable X take the values $-3, -2, -1, 0, 1, 2, 3$ such that $P(X = 0) = P(X < 0)$ and $P(X = -3) = P(X = -2) = P(X = -1) = P(X = 1) = P(X = 2) = P(X = 3)$. Find the probability distribution.

Sol: Let the distribution be as follows

X	-3	-2	-1	0	1	2	3
$P(X)$	p_1	p_2	p_3	p_4	p_5	p_6	p_7

By data $P(X = 0) = P(X < 0)$

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

$$\text{i.e., } p_4 = p_3 + p_2 + p_1 \quad \text{--- (1)}$$

Also we have

$$P(X = -3) = P(X = -2) = P(X = -1) = P(X = 1) = P(X = 2) = P(X = 3)$$

$$\text{i.e., } p_1 = p_2 = p_3 = p_5 = p_6 = p_7 \quad \text{--- (2)}$$

Further we must have

$$p_1 + p_2 + p_3 + p_4 + p_5 + p_6 + p_7 = 1 \quad \text{--- (3)}$$

Using (2) in (1), we get

$$\begin{aligned} p_4 &= p_1 + p_1 + p_1 \\ p_4 &= 3p_1 \end{aligned}$$

Using (2) in (3), we get

$$\begin{aligned} p_1 + p_1 + p_1 + p_4 + p_1 + p_1 + p_1 &= 1 \\ 6p_1 + p_4 &= 1 \\ 6p_1 + 3p_1 &= 1 && (\text{since } p_4 = 3p_1) \\ 9p_1 &= 1 \end{aligned}$$

$$p_1 = \frac{1}{9}$$

$$\text{Hence } p_4 = \frac{3}{9}$$

$$\text{Thus , } p_2 = p_3 = p_5 = p_6 = p_7 = \frac{1}{9}$$

Thus the probability distribution is as follows

X	-3	-2	-1	0	1	2	3
$P(X)$	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{3}{9}$	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$

7. The *p.d.f.* of variate X is given by the following table.

x	0	1	2	3	4	5	6
$P(x)$	K	3K	5K	7K	9K	11K	13K

For what value of K , this represents a valid probability distribution? Also find $P(x \geq 5)$ and $P(3 < x \leq 6)$.

Sol: The probability distribution is valid if $P(x) \geq 0$ and $\sum P(x) = 1$

Hence we must have $K \geq 0$ and $K + 3K + 5K + 7K + 9K + 11K + 13K = 1$



ATME

College of Engineering



$$49K = 1$$
$$K = \frac{1}{49}$$

Also, $P(x \geq 5) = P(5) + P(6)$

$$P(x \geq 5) = 11K + 13K = 24K$$

$$P(x \geq 5) = \frac{24}{49}$$

Now, $P(3 < x \leq 6) = P(4) + P(5) + P(6)$

$$P(3 < x \leq 6) = 9K + 11K + 13K$$

$$P(3 < x \leq 6) = 33K$$

$$P(3 < x \leq 6) = \frac{33}{49}$$



A T M E
College of Engineering



Module-4: Joint probability distribution

Joint Probability Distribution

If X and Y are two discrete random variables, we define the *joint probability function* of X and Y by $P(X = x, Y = y) = f(x, y)$ where $f(x, y)$ satisfy the conditions $f(x, y) \geq 0$ and $\sum_x \sum_y f(x, y) = 1$.

Suppose $X = \{x_1, x_2, \dots, x_m\}$ and $Y = \{y_1, y_2, \dots, y_n\}$ then,
 $P(X = x_i, Y = y_j) = f(x_i, y_j)$ is denoted by J_{ij} .

Expectation, Variance, Covariance and Correlation

If X is a discrete random variable taking values $x_1, x_2, \dots, x_n$ having probability function $f(x)$ then,

The *expectation* of X is denoted by $\mu_X = E(X) = \sum_{i=1}^n x_i f(x_i)$

The *Variance* of X is denoted by $V(X) = \sum_{i=1}^n x_i^2 f(x_i) - \mu^2$, where μ is the mean of X .

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

The *Standard deviation* of X is denoted by $\sigma_X = \sqrt{V(X)}$

If X and Y are two discrete random variables having the joint probability distribution $f(x, y)$ then the *expectation* of X and Y are defined as

$$\mu_X = E(X) = \sum_i x_i f(x_i)$$

$$\mu_Y = E(Y) = \sum_j y_j g(y_j)$$

and $E(XY) = \sum_{i,j} x_i y_j J_{ij}$

If X and Y are random variables having mean μ_X [or $E(X)$] and μ_Y [or $E(Y)$] respectively, then the *covariance* of X and Y are defined as

$$\text{Cov}(X, Y) = E(XY) - E(X)E(Y)$$

The **correlation** of X and Y denoted by $\rho(X, Y)$ is given by $\rho(X, Y) = \frac{Cov(X, Y)}{\sigma_X \sigma_Y}$ where
 $\sigma_X^2 = E(X^2) - \mu_X^2$, $\sigma_Y^2 = E(Y^2) - \mu_Y^2$

Note: If X and Y are independent random variables then

- i) $E(XY) = E(X)E(Y)$
- ii) $Cov(X, Y) = 0$ and $\rho(X, Y) = 0$

Problems

1. The joint distribution of two random variables X and Y is as follows

$X \downarrow \quad Y \rightarrow$	-4	2	7
1	$\frac{1}{8}$	$\frac{1}{4}$	$\frac{1}{8}$
5	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{8}$

Compute the following
 i) $E(X)$ and $E(Y)$ ii) $E(XY)$ iii) σ_X and σ_Y iv) $Cov(X,Y)$ v) $\rho(X,Y)$
Sol: The distribution (marginal distribution) of X and Y is as follows.

$X \downarrow \quad Y \rightarrow$	- 4	2	7	$f(x_i)$
1	$\frac{1}{8}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{8} + \frac{1}{4} + \frac{1}{8} = \frac{1}{2}$
5	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{4} + \frac{1}{8} + \frac{1}{8} = \frac{1}{2}$
$g(y_j)$	$\frac{1}{8} + \frac{1}{4} = \frac{3}{8}$	$\frac{1}{4} + \frac{1}{8} = \frac{3}{8}$	$\frac{1}{8} + \frac{1}{8} = \frac{1}{4}$	1

Distribution of X

x_i	1	5
$f(x_i)$	$\frac{1}{2}$	$\frac{1}{2}$

Distribution of Y

y_j	- 4	2	7
$g(y_j)$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{4}$

i) $E(X) = \sum_i x_i f(x_i)$

$$E(X) = (1) \left(\frac{1}{2}\right) + (5) \left(\frac{1}{2}\right)$$

$$E(X) = \frac{1}{2} + \frac{5}{2}$$

Thus, $E(X) = \mu_X = 3$

$$E(Y) = \sum_j y_j g(y_j)$$

$$E(Y) = (-4) \left(\frac{3}{8}\right) + (2) \left(\frac{3}{8}\right) + (7) \left(\frac{1}{4}\right)$$

$$E(Y) = -\frac{3}{2} + \frac{3}{4} + \frac{7}{4}$$

$E(Y) = \mu_Y = 1$

ii) $E(XY) = \sum_{i,j} x_i y_j J_{ij}$

$$E(XY) = (1)(-4) \left(\frac{1}{8}\right) + (1)(2) \left(\frac{1}{4}\right) + (1)(7) \left(\frac{1}{8}\right) + (5)(-4) \left(\frac{1}{4}\right) + (5)(2) \left(\frac{1}{8}\right) + (5)(7) \left(\frac{1}{8}\right)$$

$$E(XY) = -\frac{1}{4} + \frac{1}{2} + \frac{7}{8} - 5 + \frac{5}{4} + \frac{35}{8}$$

Thus, $E(XY) = \frac{3}{2}$

$$\text{iii) } \sigma_X^2 = E(X^2) - \mu_X^2 \quad ; \quad \sigma_Y^2 = E(Y^2) - \mu_Y^2$$

$$\text{Now, } E(X^2) = \sum_i x_i^2 f(x_i)$$

$$E(X^2) = (1^2) \left(\frac{1}{2}\right) + (5^2) \left(\frac{1}{2}\right)$$

$$E(X^2) = 13$$

$$E(Y^2) = \sum_j y_j^2 f(y_j)$$

$$E(Y^2) = ((-4)^2) \left(\frac{3}{8}\right) + (2^2) \left(\frac{3}{8}\right) + (7^2) \left(\frac{1}{4}\right)$$

$$E(Y^2) = \frac{79}{4}$$

$$\therefore \sigma_X^2 = 13 - 3^2 \quad ; \quad \sigma_Y^2 = \frac{79}{4} - 1^2$$

$$\sigma_X^2 = 4 \quad ; \quad \sigma_Y^2 = \frac{75}{4}$$

$$\sigma_X = 2 \quad ; \quad \sigma_Y = 4.33$$

$$\text{iv) } \text{Cov}(X, Y) = E(XY) - E(X)E(Y)$$

$$\text{Cov}(X, Y) = \frac{3}{2} - (3)(1)$$

$$\text{Cov}(X, Y) = -\frac{3}{2}$$

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

$$\rho(X, Y) = \frac{-\frac{3}{2}}{(2)(4.33)}$$

$$\rho(X, Y) = -0.1732$$

2. The joint probability distribution table for two random variables X and Y is as follows

$Y \rightarrow$ $X \downarrow$	-2	-1	4	5
1	0.1	0.2	0	0.3
2	0.2	0.1	0.1	0

Determine the marginal distributions of X and Y . Also compute i) Expectations of X, Y and XY
ii) Standard deviations of X, Y iii) Covariance of X and Y iv) correlation of X and Y

Further verify that X and Y are dependent random variables. Also find $P(X + Y > 0)$.

Sol: Marginal distribution of X

x_i	1	2
$f(x_i)$	0.6	0.4

$$i) \quad E(X) = \sum_i x_i f(x_i)$$
$$E(X) = (1)(0.6) + (2)(0.4)$$

$$E(X) = 0.6 + 0.8$$

$$\text{Thus, } E(X) = \mu_X = 1.4$$

Marginal distribution of Y

y_j	-2	-1	4	5
$g(y_j)$	0.3	0.3	0.1	0.3

$$E(Y) = \sum_j y_j g(y_j)$$
$$E(Y) = (-2)(0.3) + (-1)(0.3) + (4)(0.1) + (5)(0.3)$$

$$E(Y) = -0.6 - 0.3 + 0.4 + 1.5$$

$$E(Y) = \mu_Y = 1$$

$$E(XY) = \sum_{i,j} x_i y_j J_{ij}$$

$$\begin{aligned} E(XY) &= (1)(-2)(0.1) + (1)(-1)(0.2) + (1)(4)(0) + (1)(5)(0.3) + (2)(-2)(0.2) + (2)(-1)(0.1) + (2)(4)(0.1) \\ &\quad + (2)(5)(0) \end{aligned}$$

$$E(XY) = -0.2 - 0.2 + 0 + 1.5 - 0.8 - 0.2 + 0.8 + 0$$

Thus, $E(XY) = 0.9$

ii) $\sigma_X^2 = E(X^2) - \mu_X^2$

Now, $E(X^2) = \sum_i x_i^2 f(x_i)$

$$E(X^2) = (1^2)(0.6) + (2^2)(0.4)$$

$$E(X^2) = 2.2$$

; $\sigma_Y^2 = E(Y^2) - \mu_Y^2$

$$E(Y^2) = \sum_j y_j^2 g(y_j)$$

$$E(Y^2) = ((-2)^2)(0.3) + ((-1)^2)(0.3) + (4^2)(0.1) + (5^2)(0.3)$$

$$E(Y^2) = 10.6$$

$$\therefore \sigma_X^2 = 2.2 - (1.4)^2 \quad ; \quad \sigma_Y^2 = 10.6 - 1^2$$

$$\sigma_X^2 = 0.24 \quad ; \quad \sigma_Y^2 = 9.6$$

$$\sigma_X = 0.49 \quad ; \quad \sigma_Y = 3.1$$

$$\text{iii) } Cov(X, Y) = E(XY) - E(X)E(Y)$$

$$Cov(X, Y) = 0.9 - (1.4)(1)$$

$$Cov(X, Y) = -0.5$$

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

$$\rho(X, Y) = \frac{-0.5}{(0.49)(3.1)}$$

$$\rho(X, Y) = -0.3$$

If X and Y are independent random variables we must have $f(x_i) \cdot g(y_j) = J_{ij}$

It can be seen that $f(x_1) \cdot g(y_1) = (0.6)(0.3) = 0.18$ and $J_{11} = 0.1$ i.e., $f(x_i) \cdot g(y_j) \neq J_{ij}$

Similarly for others also the condition is not satisfied.

Hence we conclude that X and Y are dependent random variables.

We have, $X = \{x_i\} = \{x_1, x_2\} = \{1, 2\}$ respectively.

$Y = \{y_j\} = \{y_1, y_2, y_3, y_4\} = \{-2, -1, 4, 5\}$ respectively.

Also, $J_{11} = 0.1$, $J_{12} = 0.2$, $J_{13} = 0$, $J_{14} = 0.3$

$J_{21} = 0.2$, $J_{22} = 0.1$, $J_{23} = 0.1$, $J_{24} = 0$

$X + Y > 0$ is possible when (X, Y) take the values

$(x_1, y_3) = (1, 4)$; $(x_1, y_4) = (1, 5)$; $(x_2, y_2) = (2, -1)$; $(x_2, y_3) = (2, 4)$ and

$(x_2, y_4) = (2, 5)$

Hence , $P(X + Y > 0) = J_{13} + J_{14} + J_{22} + J_{23} + J_{24}$

$$P(X + Y > 0) = 0 + 0.3 + 0.1 + 0.1 + 0$$

Thus, $P(X + Y > 0) = 0.5$

3. Suppose X and Y are independent random variables with the following respective distribution, find the joint distribution of X and Y . Also verify $Cov(X, Y) = 0$.

x_i	1	2
$f(x_i)$	0.7	0.3

y_j	-2	5	8
$g(y_j)$	0.3	0.5	0.2

Sol: Since X and Y are independent, the joint distribution $J(x, y)$ is obtained by using the definition $f(x_i) \cdot g(y_j) = J_{ij}$.

J_{ij} are obtained on multiplication of the marginal entries.

From the given data , the required J_{ij} is found

$J_{11} = f(x_1).g(y_1) = (0.7)(0.3) = 0.21$
 $J_{12} = f(x_1).g(y_2) = (0.7)(0.5) = 0.35$
 $J_{13} = f(x_1).g(y_3) = (0.7)(0.2) = 0.14$
 $J_{21} = f(x_2).g(y_1) = (0.3)(0.3) = 0.09$
 $J_{22} = f(x_2).g(y_2) = (0.3)(0.5) = 0.15$
 $J_{23} = f(x_2).g(y_3) = (0.3)(0.2) = 0.06$

The joint distribution table is as follows

Y → X ↓	-2	5	8	$f(x_i)$
1	0.21	0.35	0.14	0.7
2	0.09	0.15	0.06	0.3
$g(y_j)$	0.3	0.5	0.2	1

We have $Cov(X, Y) = E(XY) - E(X)E(Y)$

$$E(X) = \sum_i x_i f(x_i)$$

$$E(X) = (1)(0.7) + (2)(0.3)$$

$$E(X) = 0.7 + 0.6$$

$$\text{Thus, } \mathbf{E(X) = \mu_X = 1.3}$$

$$E(Y) = \sum_j y_j g(y_j)$$

$$E(Y) = (-2)(0.3) + (5)(0.5) + (8)(0.2)$$

$$E(Y) = -0.6 + 2.5 + 1.6$$

$$\mathbf{E(Y) = \mu_Y = 3.5}$$

$$E(XY) = \sum_{i,j} x_i y_j J_{ij}$$

$$E(XY)$$

$$= (1)(-2)(0.21) + (1)(5)(0.35) + (1)(8)(0.14) + (2)(-2)(0.09) + (2)(5)(0.15) + (2)(8)(0.06)$$

$$E(XY) = -0.42 + 1.75 + 1.12 - 0.36 + 1.5 + 0.96$$

$$\text{Thus, } \mathbf{E(XY) = 4.55}$$



ATME

College of Engineering



Module-5: Statistical Inference



Sampling Theory

Random Sampling

A large collection individuals or attributes or numerical data can be understood as a *population* or *universe*.

A finite subset of the universe is called a *sample*. The number of individuals in a sample is called a *sample size*(n).

If $n \leq 30$ then the sample is said to be *small* otherwise it is a *large sample*.

The selection of an individual or item from the population in such a way that each has the same chance of being selected is called as *random sampling*.

Sampling where a member of the population may be selected more than once is called as *sampling with replacement*.

If a member cannot be chosen more than once is called as *sampling without replacement*.



Parameter and Statistics

The statistical constants of population is mean(μ), standard deviation (σ), correlation (ρ) are called the *parameter*.

The statistical constants drawn from the given population is mean($\bar{x}$), standard deviation (s), correlation (r) are called *statistic*.

Testing of Hypothesis

To reach the decisions about the populations and the basis of sample information, we make certain assumptions about the population involved. Such assumptions which may or may not be true are called *hypothesis*. **OR** Quantitative statement about the population. Explanation on the basis of limited evidences.

Null Hypothesis

The hypothesis formulated for the purpose of its rejection under the assumption that it is true is called the *Null Hypothesis* denoted by H_0 .

Any hypothesis which is complimentary to the null hypothesis is called *Alternative Hypothesis* denoted by H_1 .

Type-I and Type-II Errors

Type-I Error: Wrong decision to reject the null hypothesis when it is actually true.

Type-II Error: Wrong decision to accept the null hypothesis when it is actually false.

	Accepting the hypothesis	Rejecting the hypothesis
Hypothesis True	Correct decision	Wrong decision (Type-I Error)
Hypothesis False	Wrong decision (Type-II Error)	Correct decision



Significance level

The probability level, below which leads to the rejection of the hypothesis is known as the *significance level*. The probability levels are fixed at 0.05 or 0.01 being 5% or 1%. These are called *Significance level*.

Test of Significance

The process which enables us to decide about the acceptance or rejection of the hypothesis is called the *test of significance*.

Confidence Intervals

Let us suppose that we have a normal population with mean μ and S.D σ . If $\bar{x}$ is the sample mean of a random sample size n the quantity z defined by $z = \frac{\bar{x} - \mu}{\left(\frac{\sigma}{\sqrt{n}}\right)}$ is called the *Standard normal variate*.

From the normal distribution table we find the 95% of the area lies between $z = -1.96$ and $z = +1.96$ i.e., 95% confidence interval lies between -1.96 and $+1.96$.

Further 5% level of significance is denoted by $z_{0.05}$. Mathematically we can write

$$-1.96 \leq \frac{\bar{x} - \mu}{\left(\frac{\sigma}{\sqrt{n}}\right)} \leq 1.96$$

$$-\frac{\sigma}{\sqrt{n}}(1.96) \leq \bar{x} - \mu \leq (1.96)\frac{\sigma}{\sqrt{n}}$$
$$\mu \leq \bar{x} + \frac{\sigma}{\sqrt{n}}(1.96) \quad \text{and} \quad \bar{x} - \frac{\sigma}{\sqrt{n}}(1.96) \leq \mu$$

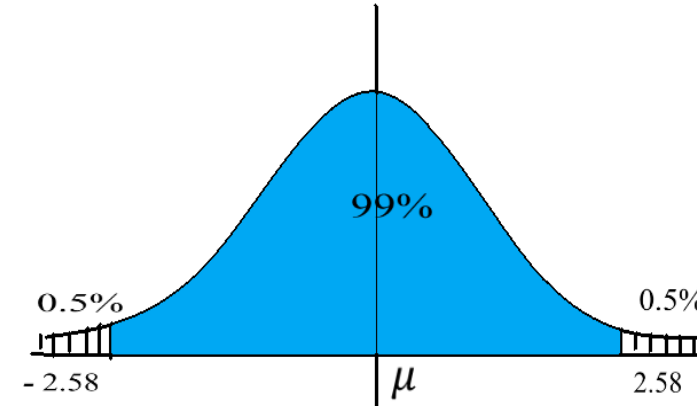
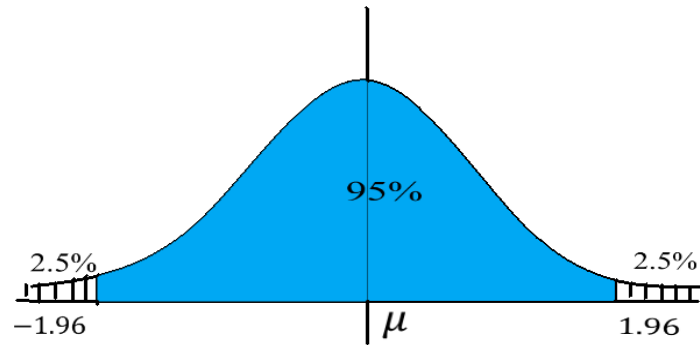
Also we can write combining the two results in the form,

$$\bar{x} - 1.96\left(\frac{\sigma}{\sqrt{n}}\right) \leq \mu \leq \bar{x} + 1.96\left(\frac{\sigma}{\sqrt{n}}\right) \quad \text{--- (1)}$$

Similarly, from the table of normal areas 99% of the area lies between -2.58 and 2.58. This is equivalent to the form,

$$\bar{x} - 2.58\left(\frac{\sigma}{\sqrt{n}}\right) \leq \mu \leq \bar{x} + 2.58\left(\frac{\sigma}{\sqrt{n}}\right) \quad \text{--- (2)}$$

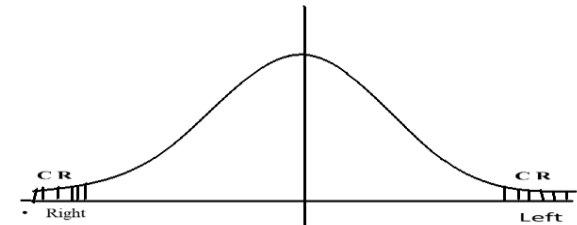
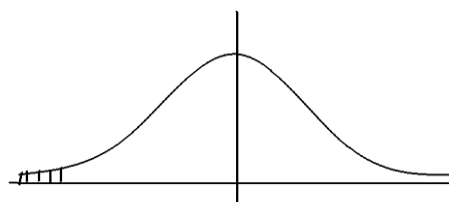
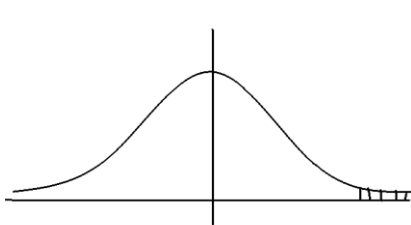
We can say that (1) is the 95% confidence interval and (2) is the 99% confidence interval.



One tailed and two tailed tests

In our test of acceptance or non acceptance of a hypothesis, we concentrated on the value of z on both sides of the mean. This can be categorically stated that, the focus of the attention lies in the two “tails” of the distribution and hence such a test is called a *two tailed test*.

Sometimes we will be interested in the extreme values to only one side of the mean in which the region of significance will be a region to one side of the distribution. Obviously the area of such a region will be equal to the level of significance itself. Such a tail is called a *one tailed test*.



Right **Critical Region** Left

One tailed test

Two tailed test

The following table will be useful for working problems

Test	Critical Value of z	
	5% level	1% level
One-tailed test	-1.645 or 1.645	-2.33 or 2.33
Two-tailed test	-1.96 or 1.96	-2.58 or 2.58

Tests of significance of Proportions

1. Let x be the observed number of successes in a sample size of n and $\mu = np$ be the expected number of successes let the associated standard normal variate Z be defined by $z = \frac{x-\mu}{\sigma} = \frac{x-np}{\sqrt{npq}}$
2. If $|Z| > 2.58$, we conclude that the differences is highly significant and reject the hypothesis .Probable limits $= p \pm 2.58 \sqrt{\frac{pq}{n}}$
3. If $Z < 1.96$, difference between the observed and expected number of successes is not significant.
4. If $Z > 1.96$, difference is significant at 5% level of significance.



ATME

College of Engineering



Test of hypothesis for means

Let μ_1 and μ_2 be the mean of two populations. Let $(\bar{x}_1, \sigma_1)$; $(\bar{x}_2, \sigma_2)$ be the mean and standard deviation of two large samples of size n_1 and n_2 respectively. To test the null hypothesis H_0 that there is no difference between the population means. i.e., $H_0: \mu_1 = \mu_2$

The static for this is given by
$$Z = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Test of significance of proportions

1. A coin is tossed 1000 times and head turns up 540 times. Decide on the hypothesis that the coin is unbiased.

Sol: Let us suppose that the coin is unbiased

p = Probability of getting a head in one toss = $\frac{1}{2}$

Since $p + q = 1$, $q = \frac{1}{2}$

Expected number of heads in 1000 tosses = np
 $= 1000 \times \frac{1}{2} = 500$

Actual Number of heads = 540 = x

Now $x - np = 540 - 500 = 40$

Consider $z = \frac{x - \mu}{\sqrt{npq}} = \frac{x - np}{\sqrt{npq}} = \frac{540 - 500}{\sqrt{1000 \times 0.5 \times 0.5}}$

$$z = 2.53 < 2.58$$

Thus, we can say that the coin is unbiased.

2. A sample of 900 days was taken in a coastal town and it was found that on 100 days the weather was very hot. Obtain the probable limits of the percentage of very hot weather.

Sol: Probability of very hot weather, $p = \frac{100}{900} = \frac{1}{9}$

$$\therefore q = \frac{8}{9}, n = 900$$

$$\begin{aligned}\text{Probable limits} &= p \pm (2.58) \sqrt{\frac{pq}{n}} \\ &= \frac{1}{9} \pm (2.58) \sqrt{\frac{1}{9} \cdot \frac{8}{9} \cdot \frac{1}{900}} \\ &= 0.111 \pm 0.027 \\ &= 0.084 \text{ and } 0.138\end{aligned}$$

Probability limits of very hot weather is 8.4% to 13.8%

3. In a sample of 500 men it was found that 60% of them had over weight. What can we infer about the proportion of people having over weight in the population?

Sol: Probability of persons having over weight is , $p = \frac{60}{100} = 0.6$

$$\therefore q = 1 - p = 0.4 \quad , n = 500$$

$$\begin{aligned}\text{Probable limits} &= p \pm (2.58) \sqrt{\frac{pq}{n}} \\ &= 0.6 \pm (2.58) \sqrt{\frac{(0.6)(0.4)}{500}} \\ &= 0.6 \pm 0.0565\end{aligned}$$

Probable limits = 0.5435 and 0.6565

Thus, the probable limits of people having over weight is 54.35% and 65.65%

4. A survey was conducted in a slum locality of 2000 families by selecting a sample size 800. It was revealed that 180 families were illiterates. Find the probable limits of the illiterate families in the population of 2000.

Sol: Probability of illiterate families is , $p = \frac{180}{800} = 0.225$

$$\therefore q = 1 - p = 0.775, n = 800$$

$$\begin{aligned}\text{Probable limits} &= p \pm (2.58) \sqrt{\frac{pq}{n}} \\ &= 0.225 \pm (2.58) \sqrt{\frac{(0.225)(0.775)}{800}} \\ &= 0.225 \pm 0.038\end{aligned}$$

$$\text{Probable limits} = 0.187 \text{ and } 0.263$$

The probable limits of the illiterate families in the population of 2000

$$= 0.187 * 2000 \text{ and } 0.263 * 2000$$

$$= 374 \text{ and } 526$$

Thus, 374 to 526 are probably illiterate families.

5. In 324 throws of a six faced 'die', an odd number turned up 181 times. Is it reasonable to think that the 'die' is an unbiased one?

Sol: Probability of the turn up of an odd number , $p = 3/6 = 1/2$

Since $p + q = 1$, $q = 1/2$

Expected number of successes = np

$$= 324 \times 1/2 = 162$$

Actual Number of heads = 181 = x

Now $x - np = 181 - 162 = 19$

Consider $z = \frac{x - \mu}{\sqrt{npq}} = \frac{x - np}{\sqrt{npq}} = \frac{19}{\sqrt{324 \times 0.5 \times 0.5}} = \frac{19}{9}$

$$z = 2.11 < 2.58$$

Thus, we can say that the die is unbiased.

6. A sample of 100 days is taken from meteorological records of a certain district and 10 of them are found to be foggy. What are the probable limits of the percentage of foggy days in the district.

Sol: Proportion of foggy days in a sample of 100 days is , $p = \frac{10}{100} = 0.1$

$$\therefore q = 1 - p = 0.9 \quad , n = 100$$

$$\begin{aligned} \text{Probable limits} &= p \pm (2.58) \sqrt{\frac{pq}{n}} \\ &= 0.1 \pm (2.58) \sqrt{\frac{(0.1)(0.9)}{100}} \\ &= 0.1 \pm 0.0774 \end{aligned}$$

Probable limits = 0.0226 and 0.1774

Thus, the percentage of foggy days lies between 2.26% and 17.74%

7. A random sample of 500 apples was taken from a large consignment and 65 were found to be bad. Estimate the proportion of bad apples in the consignment as well as the standard error of the estimate. Also find the percentage of bad apples in the consignment.

Sol: Proportion of bad apples in the sample is given by , $p = \frac{65}{500} = 0.13$

$$\therefore q = 1 - p = 0.87 \quad , n = 500$$

$$\text{S.E proportion of bad apples} = \sqrt{\frac{pq}{n}} = \sqrt{\frac{(0.13)(0.87)}{500}} = 0.015$$

$$\begin{aligned} \text{Probable limits} &= p \pm (2.58) \sqrt{\frac{pq}{n}} \\ &= 0.13 \pm (2.58)(0.015) \\ &= 0.13 \pm 0.0387 \end{aligned}$$

Probable limits = 0.0913 and 0.1687
= 9.13% and 16.87%

Thus, the required percentage of bad apples in the consignment lies between 9.13 and 16.87

8. The mean and standard deviation of the maximum loads supported by 60 cables are 11.09 tonnes and 0.73 tonnes respectively. Find a) 95% b) 99% confidence limits for mean of the maximum loads of all cables produced by the company.

Sol: By data $\bar{x} = 11.09$, $\sigma = 0.73$, $n = 60$

a) 95% confidence limits for mean of the maximum loads are given by

$$\begin{aligned}\bar{x} \pm 1.96 \left(\frac{\sigma}{\sqrt{n}} \right) &= 11.09 \pm (1.96) \left(\frac{0.73}{\sqrt{60}} \right) \\ &= 11.09 \pm 0.18 \\ &= 10.91 \text{ and } 11.27\end{aligned}$$

Thus, 10.91 tonnes to 11.27 tonnes are the 95% confidence limits for the mean of the maximum loads.

b) 99% confidence limits for mean of the maximum loads are given by

$$\begin{aligned}\bar{x} \pm 2.58 \left(\frac{\sigma}{\sqrt{n}} \right) &= 11.09 \pm (2.58) \left(\frac{0.73}{\sqrt{60}} \right) \\ &= 11.09 \pm 0.24 \\ &= 10.85 \text{ and } 11.33\end{aligned}$$

Thus, 10.85 tonnes to 11.33 tonnes are the 95% confidence limits for the mean of the maximum loads.

9. An unbiased coin is thrown n times. It is desired that the relative frequency of the appearance of heads should lie between 0.49 and 0.51. Find the smallest value of n that will ensure this result with a) 95% confidence b) 99% confidence

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

$$q = \frac{1}{2}$$

$$\text{S.E proportion of heads} = \sqrt{\frac{pq}{n}}$$

$$\text{S.E proportion of heads} = \sqrt{\frac{\frac{1}{2} \cdot \frac{1}{2}}{n}} = \frac{1}{2\sqrt{n}}$$

a) Probable limits for 95% confidence level is given by : $p \pm (1.96) \sqrt{\frac{pq}{n}}$, which should lie between 0.51 and 0.49

$$p \pm (1.96) \sqrt{\frac{pq}{n}} = 0.51 \text{ or } 0.49$$

$$0.5 \pm (1.96) \frac{1}{2\sqrt{n}} = 0.51 \text{ or } 0.49$$

$$0.5 \pm \frac{1.96}{2\sqrt{n}} = 0.51 \text{ or } 0.49$$

$$0.5 + \frac{1.96}{2\sqrt{n}} = 0.51$$

$$0.5 - \frac{1.96}{2\sqrt{n}} = 0.49$$

$$\frac{1.96}{2\sqrt{n}} = 0.51 - 0.5$$

$$\frac{1.96}{2\sqrt{n}} = 0.5 - 0.49$$

$$\frac{1.96}{2\sqrt{n}} = 0.01$$

$$\frac{1.96}{2\sqrt{n}} = 0.01$$

$$\sqrt{n} = \frac{1.96}{0.02}$$

$$\sqrt{n} = 98$$

Thus, **n = 9604**

b) Taking the confidence coefficient equal to 1.645 for 90% confidence level we have as

before

$$\frac{1.645}{2\sqrt{n}} = 0.01$$



ATME

College of Engineering



$$\sqrt{n} = \frac{1.645}{0.02}$$

$$\sqrt{n} = 82.25$$

Thus, **$n = 6765$**



Test of significance of a sample mean

1. A manufacturer claimed that atleast 95% of the equipment which he supplied to a factory conformed to specifications. An examination of a sample of 200 pieces of equipment revealed that 18 of them were faulty. Test his claim at a significance level of 1% and 5%.

Sol: Let p be the probability of success which being the probability of the equipment supplied to the factory conformal to the specifications.

$\therefore$ By data $p = 0.95$ and hence $q = 0.05$

$H_0: p = 0.95$ and the claim is correct.

$H_1: p < 0.95$ and the claim is false.

We choose the one tailed test to determine whether the supply is conformal to the specification.

$$\mu = np = 200 * 0.95 = 190$$

$$\sigma = \sqrt{npq} = \sqrt{200 * 0.95 * 0.05} = 3.082$$

Expected number of equipment's according to the specification =190

Actual number =182 , since 18 out of 200 were faulty.

$$\text{Now, } Z = \frac{x - np}{\sigma}$$

$$Z = \frac{190 - 182}{3.082}$$

$$Z = 2.6$$

The value of Z is greater than the critical value 1.645 at 5% level and 2.33 at 1% level of significance.

The claim of the manufacturer (*null hypothesis that claim is correct*) is rejected at 5% as well as 1% level of significance in accordance with the one tailed test.

2. It has been found from experience that the mean breaking strength of a particular brand of thread is 275.6gms with standard deviation of 39.7gms. Recently a sample of 36 pieces of thread showed a mean breaking strength of 253.2gms. Can one conclude at a significance level of a) 0.05 b) 0.01 that the thread has become inferior ?

Sol: We have to decide between the two hypothesis,

$H_0: \mu = 275.6 \text{ gms}$, mean breaking strength.

$H_1: \mu < 275.6 \text{ gms}$, inferior in breaking strength.

We choose the one tailed test.

Mean breaking strength of a sample of 36 pieces=253.2

$\therefore$ Difference =275.6 – 253.2=22.4 ; n = 36

$$\text{Now, } Z = \frac{\text{Difference}}{(\sigma/\sqrt{n})}$$

$$Z = \frac{22.4}{(39.7/\sqrt{36})}$$
$$Z = 3.38$$

The value of Z is greater than the critical value of Z=1.645 at 5% level and 2.33 at 1% level of significance.

Under the hypothesis H_1 that the thread has become inferior is accepted at both 0.05 and 0.01 levels in accordance with one tailed test.

Test of significance of difference between means

1. In an elementary school examination the mean grade of 32 boys was 72 with a standard deviation of 8, while the mean grade of 36 girls was 75 with a standard deviation of 6.

Test the hypothesis that the performance of girls is better than boys.

Sol: We have , $\bar{x}_B = 72$, $\sigma_B = 8$, $n_B = 32$ [Boys]

$\bar{x}_G = 75$, $\sigma_G = 6$, $n_G = 36$ [Girls]

Consider, $Z = \frac{\bar{x}_G - \bar{x}_B}{\sqrt{\frac{\sigma_G^2}{n_G} + \frac{\sigma_B^2}{n_B}}}$

$$Z = \frac{75 - 72}{\sqrt{\frac{36}{36} + \frac{64}{32}}}$$

$$Z = \sqrt{3}, Z = 1.73$$

$$Z = 1.73 \begin{cases} > Z_{0.05} = 1.645 \text{ (one tailed test)} \\ < Z_{0.01} = 2.33 \text{ (one tailed test)} \end{cases}$$

The difference in the performance of girls and boys in the examination is significant at 5% level but not at 1% level.

2. A sample of 100 bulbs produced by a company A showed a mean life of 1190 hours and a standard deviation of 90 hours. Also a sample of 75 bulbs produced by a company B showed a mean life of 1230 hours and a standard deviation of 120 hours. Is there a difference between the mean life time of the bulbs produced by the two companies at
a) 5% level of significance b) 1% level of significance.

Sol: By data $\bar{x}_A = 1190$, $\sigma_A = 90$, $n_A = 100$ [Company A]

$\bar{x}_B = 1230$, $\sigma_B = 120$, $n_B = 75$ [Company B]

Consider, $Z = \frac{\bar{x}_B - \bar{x}_A}{\sqrt{\frac{\sigma_B^2}{n_B} + \frac{\sigma_A^2}{n_A}}}$

$$Z = \frac{1230 - 1190}{\sqrt{\frac{(120)^2}{75} + \frac{(90)^2}{100}}}$$

$$Z = 2.42$$

$$Z = 2.42 \begin{cases} > Z_{0.05} = 1.96 \text{ (Two tailed test)} \\ < Z_{0.01} = 2.58 \text{ (Two tailed test)} \end{cases}$$

The null hypothesis that there is no difference between the mean life time of bulbs is rejected at 5% level but not at 1% level of significance.

The null hypothesis that there is no difference between the mean life time of bulbs is rejected at both levels of significance in a one tailed test as the respective critical values are 1.645 and 2.33.

3. The mean life of two large samples of 1000 and 2000 members are 168.75 cms and 170 cms respectively. Can this be regarded as drawn from the sample population of standard deviation 6.25 cms ?

Sol: By data $\bar{x}_1 = 168.75$, $n_1 = 1000$

$\bar{x}_2 = 170$, $n_2 = 2000$, $\sigma = 6.25$

Consider, $Z = \frac{\bar{x}_2 - \bar{x}_1}{\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$

$$Z = \frac{170 - 168.75}{6.25 \sqrt{\frac{1}{1000} + \frac{1}{2000}}}$$
$$Z = 5.16$$

$Z = 5.16$ is very much greater than $Z_{0.05} = 1.96$ and $Z_{0.01} = 2.58$. Thus we say that the difference between the sample means is significant and we conclude that the samples cannot be regarded as drawn from the same population.

4. A random sample for 1000 workers in company has mean wage of Rs.50 per day and S.D of Rs.15. Another sample of 1500 workers from another company has mean wage of Rs.45 per day and S.D of Rs.20. Does the mean rate of wages varies between the two companies? Find the 95% confidence limits for the difference of the wages of the population of the two companies?

Sol: By data $\bar{x}_1 = 50$, $\sigma_1 = 15$, $n_1 = 1000$ [Company 1]
 $\bar{x}_2 = 45$, $\sigma_2 = 20$, $n_2 = 1500$ [Company 2]

Consider,
$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

$$Z = \frac{50 - 45}{\sqrt{\frac{15^2}{1000} + \frac{20^2}{1500}}} \quad , Z = 7.1307$$

The value of $Z = 7.1307$ is greater than $Z_{0.05} = 1.96$ and $Z_{0.01} = 2.58$. Hence we can say that the difference between the mean wages is significant both at 5% and 1% level of significance.

Also 95% confidence limits for the difference of mean wages is given by

$$\begin{aligned}(\bar{x}_1 - \bar{x}_2) \pm 1.96 \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} &= 5 \pm 1.96(0.7012) \\&= 5 \pm 1.374 \\&= 3.626 \text{ and } 6.374\end{aligned}$$

Thus we can say with 95% confidence that the difference of population mean of wages between the two companies lies between Rs.3.63 and Rs.6.37

Student's t distribution/test

We need to test the hypothesis, whether the sample mean $\bar{x}$ differs significantly from the population mean or hypothetical value μ .

We compute, $t = \frac{\bar{x} - \mu}{s} \sqrt{n}$ and consider $|t|$.

Here, $\bar{x} = \frac{\sum x}{n}$ and $s^2 = \frac{1}{(n-1)} \sum_{i=1}^n (x_i - \bar{x})^2$

$v = n - 1$, denote the number of degrees of freedom 't'.

If $|t| > t_{0.05}$, the difference between $\bar{x}$ and μ is said to be significant at 5% level of significance.

If $|t| > t_{0.01}$, the difference between $\bar{x}$ and μ is said to be significant at 1% level of significance.

If $|t|$ is less than the table value at a certain level of significance, the data is said to be conformal/consistent with the hypothesis that μ is the mean of the population.

We have 95% confidence limits for μ is given by $\bar{x} \pm \frac{s}{\sqrt{n}} t_{0.05}$

Similarly 99% confidence limits for μ is given by $\bar{x} \pm \frac{s}{\sqrt{n}} t_{0.01}$

Note: 1. Confidence limits are also called **Fiducial limits**.

2. Test of significance for difference between sample means is

$$t = \frac{\bar{x} - \bar{y}}{s \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Where, $s^2 = \frac{1}{n_1 + n_2 - 2} \left\{ \sum_{i=1}^{n_1} (x_i - \bar{x})^2 + \sum_{j=1}^{n_2} (y_j - \bar{y})^2 \right\}$

And degrees of freedom, $\nu = n_1 + n_2 - 2$

Problems

1. Find the student's t for the following variable values in a sample of eight:
-4, -2, -2, 0, 2, 2, 3, 3 taking the mean of the universe to be zero.

Sol: Wkt $t = \frac{\bar{x} - \mu}{s} \sqrt{n}$

By data $\mu = 0$ and we have $n = 8$

$$\bar{x} = \frac{\sum x}{n}$$

$$\bar{x} = \frac{1}{8}(-4 - 2 - 2 + 0 + 2 + 2 + 3 + 3)$$

$$\bar{x} = \frac{1}{4}$$

$$\bar{x} = 0.25$$



ATME

College of Engineering



$$s^2 = \frac{1}{(n-1)} \sum_{i=1}^n (x_i - \bar{x})^2$$

$$s^2 = \frac{1}{(8-1)} \{(-4.25)^2 + (-2.25)^2 + (-2.25)^2 + (-0.25)^2 + (1.75)^2 + (1.75)^2 + (2.75)^2 + (2.75)^2\}$$

$$s^2 = \frac{1}{(7)} (49.5)$$

$$s^2 = 7.07$$

$$\therefore s = 2.66$$

$$\text{Thus, } t = \frac{0.25-0}{2.66} \sqrt{8}$$

$$t = 0.266$$

2. A machine is expected to produce nails of length 3 inches. A random sample of 25 nails gave an average length of 3.1 inch with standard deviation 0.3. Can it be said that the machine is producing nails as per specification ? ($t_{0.05}$ for 24 d. f is 2.064)

Sol: By data , we have $\mu = 3$, $\bar{x} = 3.1$, $n = 25$, $s = 0.3$

$$t = \frac{\bar{x} - \mu}{s} \sqrt{n}$$

$$t = \frac{3.1 - 3}{0.3} \sqrt{25}$$

$$t = 1.67 < 2.064$$

Thus, the hypothesis that the machine is producing nails as per specification is accepted at 5% level of significance.

3. Ten individuals are chosen at random from a population and their heights in inches are found to be 63, 63, 66, 67, 68, 69, 70, 70, 71, 71. Test the hypothesis that the mean height of the universe is 66 inches. ($t_{0.05} = 2.262$ for 9 d. f)



Sol: We have $\mu = 66, n = 10$

$$\bar{x} = \frac{\sum x}{n}$$

$$\bar{x} = \frac{678}{10}$$

$$\bar{x} = \mathbf{67.8}$$

$$s^2 = \frac{1}{(n-1)} \sum_{i=1}^n (x_i - \bar{x})^2$$

$$s^2 = \frac{1}{9} [(63 - 67.8)^2 + (63 - 67.8)^2 + \dots + (71 - 67.8)^2]$$

$$s^2 = 9.067$$

$$s = \mathbf{3.011}$$

Wkt $t = \frac{\bar{x} - \mu}{s} \sqrt{n}$

$$t = \frac{(67.8 - 66)}{3.011} \sqrt{10}$$

$$t = 1.89 < 2.262$$

Thus the hypothesis is accepted at 5% level of significance.

4. A sample of 10 measurements of the diameter of a sphere gave a mean of 12 cm and a standard deviation 0.15 cm. Find the 95% confidence limits for the actual diameter.

Sol: By data, $n = 10$, $\bar{x} = 12$, $s = 0.15$

Also $t_{0.05} = 2.262$ for 9 d.f

Confidence limits for the actual diameter is given by $\bar{x} \pm \left[\frac{s}{\sqrt{n}} \right] t_{0.05}$

$$\bar{x} \pm \left[\frac{s}{\sqrt{n}} \right] t_{0.05} = 12 + \left[\frac{0.15}{\sqrt{10}} \right] (2.262) = 12 \pm 0.1073$$

Thus 11.893 cm to 12.107 cm is the confidence limits for the actual diameter.

5. A certain stimulus administered to each of the 12 patients resulted in the following change in blood pressure 5, 2, 8, -1, 3, 0, 6, -2, 1, 5, 0, 4. Can it be concluded that the stimulus will increase the blood pressure ? ($t_{0.05}$ for 11 d.f is 2.201)

Sol: $\bar{x} = \frac{\sum x}{n}$

$$\bar{x} = \frac{31}{12}$$

$$\bar{x} = 2.5833$$

$$s^2 = \frac{1}{(n-1)} \sum_{i=1}^n (x_i - \bar{x})^2$$

$$s^2 = \frac{1}{11} [(5 - 2.58)^2 + (2 - 2.58)^2 + \dots + (4 - 2.58)^2]$$

$$s^2 = 9.538$$

$$s = \mathbf{3.088}$$

$$\text{Wkt } t = \frac{\bar{x} - \mu}{s} \sqrt{n}$$

Let us suppose that the stimulus administration is not accompanied with increase in blood pressure, we can take $\mu = 0$.

$$t = \frac{(2.5833 - 0)}{3.088} \sqrt{12}$$

$$t = \mathbf{2.8979 > 2.201}$$

Hence the hypothesis is rejected at 5% level of significance. We conclude with 95% confidence that the stimulus in general is accompanied with increase in blood pressure.

6. A group of boys and girls were given an intelligence test. The mean score, S.D score and numbers in each group are as follows.

	Boys	Girls
Mean	74	70
S D	8	10
n	12	10

Is the difference between the means of the two groups significant at 5% level of significance ($t_{0.05}$ for 20 d. f is 2.086)

Sol: By data , $\bar{x} = 74, s_1 = 8, n_1 = 12$ [Boys]
 $\bar{y} = 70, s_2 = 10, n_2 = 10$ [Girls]

We have $t = \frac{\bar{x} - \bar{y}}{s \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$

Where, $s^2 = \frac{1}{n_1 + n_2 - 2} \left\{ \sum_{i=1}^{n_1} (x_i - \bar{x})^2 + \sum_{j=1}^{n_2} (y_j - \bar{y})^2 \right\}$

Or $s^2 = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}$

Now $s^2 = \frac{12(64) + 10(100)}{12 + 10 - 2}$

$$s^2 = \frac{1768}{20}$$

$$s^2 = 88.4$$

$$s = 9.402$$

$$\text{Hence } t = \frac{74-70}{(9.402)\sqrt{\frac{1}{12}+\frac{1}{10}}}$$

$$t = 0.9936 < 2.086$$

Thus the hypothesis that there is a difference between the means of the two groups is accepted at 5% level of significance.

7. A sample of 11 rats from a central population had an average blood viscosity of 3.92 with standard deviation of 0.61. On the basis of this sample, establish 95% fiducial limits for μ , the mean blood viscosity of the central population. ($t_{0.05} = 2.228$ for 10 d.f)

Sol: By data $\bar{x} = 3.92$, $s = 0.61$, $n_1 = 11$

95% fiducial limits for $\mu = \bar{x} \pm \left[\frac{s}{\sqrt{n}} \right] t_{0.05}$

$$\begin{aligned}\bar{x} \pm \left[\frac{s}{\sqrt{n}} \right] t_{0.05} &= 3.92 \pm \left[\frac{0.61}{\sqrt{11}} \right] (2.228) \\ &= 3.92 \pm 0.41 \\ &= 3.51 \text{ and } 4.33\end{aligned}$$

Thus 95% confidence limits for μ are 3.51 and 4.33.

8. Two types of batteries are tested for their length of life and the following results were obtained.

Battery A : $n_1 = 10$, $\bar{x}_1 = 500 \text{ hrs}$, $s_1^2 = 100$

Battery B : $n_2 = 10$, $\bar{x}_2 = 500 \text{ hrs}$, $s_1^2 = 121$

Compute student's t and test whether there is a significant difference in the two means.

Sol: By data Battery A : $n_1 = 10$, $\bar{x}_1 = 500 \text{ hrs}$, $s_1^2 = 100$

Battery B : $n_2 = 10$, $\bar{x}_2 = 560 \text{ hrs}$, $s_1^2 = 121$

$$\text{Now, } s^2 = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}$$

$$s^2 = \frac{(10 \cdot 100) + (10 \cdot 121)}{10 + 10 - 2}$$

$$s^2 = 122.78$$

$$s = 11.0806$$

$$\begin{aligned} \text{We have } t &= \frac{\bar{x}_2 - \bar{x}_1}{s \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \\ t &= \frac{60}{(11.0806) \sqrt{\frac{1}{10} + \frac{1}{10}}} \end{aligned}$$

$$t = 12.1081 > 2.101 \quad (t_{0.05} = 2.101 \text{ for } 18 \text{ d.f.})$$

This value of t is greater than the table value of t for 18 d.f at all levels of significance.

The null hypothesis that there is no significant difference in the two means rejected at all significance levels.

9. A group of 10 boys fed on a diet A and another group of 8 boys fed on a different diet B for a period of 6 months recorded the following increase in weight (lbs.)

Diet A : 5 6 8 1 12 4 3 9 6 10

Diet B : 2 3 6 8 10 1 2 8

Test whether diets A and B differ significantly regarding their effect on increase in weight.

Sol: Let the variable x correspond to the diet A and y to the diet B.

$$\bar{x} = \frac{\sum x}{n_1}$$

$$\bar{x} = \frac{64}{10}$$

$$\bar{x} = 6.4$$

$$\bar{y} = \frac{\sum y}{n_2}$$

$$\bar{y} = \frac{40}{8}$$

$$\bar{y} = 5$$

$$\sum_{i=1}^{n_1} (x_i - \bar{x})^2 = 102.4 \quad ; \quad \sum_{j=1}^{n_2} (y_j - \bar{y})^2 = 82$$

$$s^2 = \frac{1}{n_1 + n_2 - 2} \left\{ \sum_{i=1}^{n_1} (x_i - \bar{x})^2 + \sum_{j=1}^{n_2} (y_j - \bar{y})^2 \right\}$$

$$s^2 = \frac{1}{16} \{102.4 + 82\}$$

$$s^2 = \frac{184.4}{16} = 11.525$$

$$s = 3.395$$

We have $t = \frac{\bar{x} - \bar{y}}{s \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$

$$t = \frac{1.4}{(3.395)\sqrt{\frac{1}{10} + \frac{1}{8}}}$$

$$t = 0.869 < 2.12 \quad \text{for 16 d.f}$$

Thus we conclude that the two diets do not differ significantly regarding their effect on increase in weight.

10. Two horses A and B were tested according to the time (in seconds) to run a particular race with the following results.

Horse A : 28 30 32 33 33 29 34

Horse B : 29 30 30 24 27 29

Test whether you can discriminate between the two horses.

Sol: Let the variables x and y respectively correspond to horse A and horse B.

$$\bar{x} = \frac{\sum x}{n_1} = \frac{219}{7} = 31.3$$

$$\bar{y} = \frac{\sum y}{n_2} = \frac{169}{6} = 28.2$$

$$\sum_{i=1}^{n_1} (x_i - \bar{x})^2 = 31.43$$

$$\sum_{j=1}^{n_2} (y_j - \bar{y})^2 = 26.84$$

$$s^2 = \frac{1}{n_1+n_2-2} \left\{ \sum_{i=1}^{n_1} (x_i - \bar{x})^2 + \sum_{j=1}^{n_2} (y_j - \bar{y})^2 \right\}$$

$$s^2 = \frac{1}{11} (31.43 + 26.84)$$

$$s^2 = 5.2973$$

$$\therefore s = 2.3016$$

We have $t = \frac{\bar{x} - \bar{y}}{s \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$

$$t = \frac{31.3 - 28.2}{(2.3016) \sqrt{\frac{1}{7} + \frac{1}{6}}}$$

$$t = 2.42 = \begin{cases} > t_{0.05} = 2.2 \\ < t_{0.02} = 2.72 \end{cases}$$

The discrimination between the horses is significant at 5% level but not at 2% level of significance.

Chi-Square distribution

Chi-square distribution provides a measure of correspondence between the theoretical frequencies and observed frequencies.

If $O_i (i = 1, 2, \dots, n)$ and $E_i (i = 1, 2, \dots, n)$ respectively denotes a set of observed and estimated frequencies, the quantity **chi-square** denoted by χ^2 is defined as follows.

$$\chi^2 = \sum_{i=1}^n \frac{(O_i - E_i)^2}{E_i} ; \text{ degrees of freedom} = n - 1$$

Note: If the expected frequencies are less than 10, we group them suitably for computing the value of chi. Square.

Chi-square test as a test of goodness of fit

It is possible to test the hypothesis about the association of two attributes. We already discussed the fitting of Binomial distribution, Normal distribution, Poisson distribution to a given data. It is easily possible to find the theoretical frequencies from the distribution of fit.

Chi-square test helps us to test the goodness of fit of these distributions.

If the calculated value of χ^2 is less than the value of χ^2 at a specified level of significance, the hypothesis is accepted. Otherwise the hypothesis is rejected.

Problems

1. A dice is thrown 264 times and the number appearing on the face (x) follows the following frequency distribution.

x	1	2	3	4	5	6
f	40	32	28	58	54	60

Calculate the value of χ^2 .

Sol: The frequencies in the given data are the observed frequencies. Assuming that the dice is unbiased, the expected number of frequencies for the numbers 1, 2, 3, 4, 5, 6 to appear on the face is $\frac{264}{6} = 44$ each.

Now the data is as follows:

No. on the dice	1	2	3	4	5	6
Observed frequency (O_i)	40	32	28	58	54	60
Expected frequency (E_i)	44	44	44	44	44	44

$$\chi^2 = \sum_{i=1}^n \frac{(O_i - E_i)^2}{E_i}$$

$$\chi^2 = \frac{(40 - 44)^2}{44} + \frac{(32 - 44)^2}{44} + \dots + \frac{(60 - 44)^2}{44}$$

$$\chi^2 = \frac{1}{44} [16 + 144 + 256 + 196 + 100 + 256]$$

Thus, $\chi^2 = 22$

2. Five dice were thrown 96 times and the numbers 1, 2 or 3 appearing on the face of the dice follows the frequency distribution as below.

No. of dice showing 1,2 or 3	5	4	3	2	1	0
Frequency	7	19	35	24	8	3

Test the hypothesis that the data follows a binomial distribution. ($\chi^2_{0.05} = 11.07$ for 5 d.f)

Sol: The data gives the observed frequencies and we need to calculate the expected frequencies.

Probability of a single dice throwing 1, 2 or 3 is $p = \frac{3}{6} = \frac{1}{2}$

$$\therefore q = 1 - p = \frac{1}{2}$$

From the binomial distribution, $P(x) = {}^nC_x p^x q^{(n-x)}$

The theoretical frequencies of getting 5, 4, 3, 2, 1, 0 successes with 5 dice are respectively the successive terms of the binomial expansion.

They are respectively $96 * {}^5C_0 * \frac{1}{2^5}$, $96 * {}^5C_1 * \frac{1}{2^5}$, $96 * {}^5C_2 * \frac{1}{2^5}$, $96 * {}^5C_3 * \frac{1}{2^5}$, $96 * {}^5C_4 * \frac{1}{2^5}$, $96 * {}^5C_5 * \frac{1}{2^5}$ and they are 3,15,30,30,15,3.

We have the table of observed and expected frequencies.

O_i	7	19	35	24	8	3
E_i	3	15	30	30	15	3

$$\chi^2 = \sum_{i=1}^n \frac{(O_i - E_i)^2}{E_i}$$

$$\chi^2 = \frac{16}{3} + \frac{16}{15} + \frac{25}{30} + \frac{36}{30} + \frac{49}{15} + \frac{0}{3}$$

$$\chi^2 = 11.7 > \chi_{0.05}^2 = 11.07$$

Thus the hypothesis that the data follows a binomial distribution is rejected.

3. A sample analysis of examination results of 500 students was made. It was found that 220 students had failed, 170 had secured third class 90 had secured second class and 20 had secured first class. Do these figures support the general examination result which is in the ratio 4:3:2:1 for the respective categories ($\chi_{0.05}^2 = 7.81$ for 3 d.f.).

Sol: Let us take the hypothesis that these figures support to the general result in the ratio 4:3:2:1

The expected frequencies in the respective category are

$$\frac{4}{10} * 500, \frac{3}{10} * 500, \frac{2}{10} * 500, \frac{1}{10} * 500 \text{ or } 200, 150, 100, 50$$



ATME

College of Engineering

We have the following table.



O_i	220	170	90	20
E_i	200	150	100	50

$$\chi^2 = \sum_{i=1}^n \frac{(O_i - E_i)^2}{E_i}$$

$$\chi^2 = \frac{400}{200} + \frac{400}{150} + \frac{100}{100} + \frac{900}{50}$$

$$\chi^2 = 23.67 > \chi_{0.05}^2 = 7.81$$

Thus the hypothesis is rejected.

4. 4 coins are tossed 100 times and the following results were obtained. Fit a binomial distribution for the data and calculate the theoretical frequencies ($\chi^2_{0.05} = 9.49$ for 4 d.f.).

Number of heads	0	1	2	3	4
Frequency	5	29	36	25	5

Sol: Let x denote the number of heads and f the corresponding frequency. Since the data is in the form of a frequency distribution we shall first calculate the mean.

$$\text{Mean, } \mu = \frac{\sum fx}{\sum f} = \frac{(0)(5) + (1)(29) + (2)(36) + (3)(25) + (4)(5)}{5 + 29 + 36 + 25 + 5} = \frac{196}{100} = 1.96$$

From binomial distribution, $\mu = np$. Here $n=4$

$$\text{Hence } 4p = 1.96$$

$$p = 0.49$$

$$q = 0.51$$

From binomial distribution,

$$P(x) = {}^n C_x p^x q^{n-x}$$

$$P(x) = {}^4 C_x (0.49)^x (0.51)^{4-x}$$

Since 4 coins were tossed, expected(theoretical) frequencies are obtained from

$$F(x) = 100P(x)$$

$$F(x) = 100.4c_x (0.49)^x (0.51)^{4-x} \text{ where } x = 0,1,2,3,4.$$

$$F(0) = 100.4c_0 (0.49)^0 (0.51)^4 = 6.765 \cong 7$$

$$F(1) = 100.4c_1 (0.49)^1 (0.51)^3 = 25.999 \cong 26$$

$$F(2) = 100.4c_2 (0.49)^2 (0.51)^2 = 37.47 \cong 37$$

$$F(3) = 100.4c_3 (0.49)^3 (0.51)^1 = 24.0004 \cong 24$$

$$F(4) = 100.4c_4 (0.49)^4 (0.51)^0 = 5.765 \cong 6$$

Thus the required theoretical frequencies are **7,26,37,24,6**.

We have the following table

O_i	5	29	36	25	5
E_i	7	26	37	24	6

$$\chi^2 = \sum_{i=1}^n \frac{(O_i - E_i)^2}{E_i}$$

$$\chi^2 = \frac{4}{7} + \frac{9}{26} + \frac{1}{37} + \frac{1}{24} + \frac{1}{6}$$

$$\chi^2 = 1.15 < \chi_{0.05}^2 = 9.49$$

Thus the hypothesis that the fitness is good can be accepted.

5. Fit a poisson distribution for the following data and test the goodness of fit given that $\chi^2_{0.05} = 7.815$ for 3 d.f

x	0	1	2	3	4
f	122	60	15	2	1

Sol: Let x denote the number of heads and f the corresponding frequency. Since the data is in the form of a frequency distribution we shall first calculate the mean.

$$\text{Mean, } \mu = \frac{\sum fx}{\sum f} = \frac{(0)(122)+(1)(60)+(2)(15)+(3)(2)+(4)(1)}{122+60+15+2+1} = \frac{100}{200} = 0.2$$

From poisson distribution, $\mu = m$.

$$\therefore m = 0.2$$

$$\text{Wkt, } P(x) = \frac{m^x e^{-m}}{x!}$$

Let $F(x) = 200 * P(x)$

$$F(x) = 200 * \frac{m^x e^{-m}}{x!}$$

$$F(x) = 200 * \frac{(0.5)^x e^{-(0.5)}}{x!} \quad \text{But } e^{-(0.5)} = 0.6065$$

$$F(x) = 121.3 \frac{(0.5)^x}{x!}$$

$$F(0) = 121, F(1) = 61, F(2) = 15, F(3) = 3, F(4) = 0$$

Thus the required theoretical frequencies are 121,61,15,3,0.

Since the last of the expected frequency is 0 we shall club it with the previous one.

We have the following table

O_i	122	60	15	2+1=3
E_i	121	61	15	3+0=3

$$\chi^2 = \sum_{i=1}^n \frac{(O_i - E_i)^2}{E_i}$$

$$\chi^2 = \frac{1}{121} + \frac{1}{61} + 0 + 0$$

$$\chi^2 = 0.025 < \chi_{0.05}^2 = 7.815$$

Thus the hypothesis that the fitness is good can be accepted.

6. The number of accidents per day (x) as recorded in a textile industry over a period of 400 days is given below. Test the goodness of fit in respect of Poisson distribution of fit to the given data ($\chi_{0.05}^2 = 9.49$ for 4 d.f.).

x	0	1	2	3	4	5
f	173	168	37	18	3	1

Sol: Let x denote the number of heads and f the corresponding frequency. Since the data is in the form of a frequency distribution we shall first calculate the mean.

$$\text{Mean, } \mu = \frac{\sum fx}{\sum f} = \frac{(0)(173) + (1)(168) + (2)(37) + (3)(18) + (4)(3) + (5)(1)}{173 + 168 + 37 + 18 + 3 + 1} = \frac{313}{400} = 0.7825$$

From poisson distribution, $\mu = m$.

$$\therefore m = 0.7825$$

$$\text{Wkt, } P(x) = \frac{m^x e^{-m}}{x!}$$

$$\text{Let } F(x) = 400 * P(x)$$

$$F(x) = 400 * \frac{m^x e^{-m}}{x!}$$

$$F(x) = 400 * \frac{(0.7825)^x e^{-(0.7825)}}{x!} \quad \text{But } e^{-(0.7825)} = 0.4573$$

$$F(x) = 182.92 \frac{(0.7825)^x}{x!}$$

$F(0) = 183, F(1) = 143, F(2) = 56, F(3) = 15, F(4) = 3, F(5) = 0$

Thus the required theoretical frequencies are 183,143,56,15,3,0.

Since the last of the expected frequency is 0 we shall club it with the previous one.

We have the following table

O_i	173	168	37	18	3+1=4
E_i	183	143	56	15	3+0=3

$$\chi^2 = \sum_{i=1}^n \frac{(O_i - E_i)^2}{E_i}$$

$$\chi^2 = \frac{100}{183} + \frac{625}{143} + \frac{361}{56} + \frac{9}{15} + \frac{1}{3} \quad , \chi^2 = 12.297 > \chi_{0.05}^2 = 9.49$$

Thus the hypothesis that the fitness is good can be rejected.



ATME

College of Engineering



Thank You