# UNIT–I

## Basic Number Theory

The theory of numbers is the branch of mathematics which deals with properties of the whole numbers (counting numbers also called positive integers).

Euclid also made an important contribution by finding all the perfect numbers.

e.g. $6 = 1 + 2 + 3$ is a perfect number (the sum of all its proper divisors).

**Well Ordering Principle: If A is a non-empty set of positive integers, then A contains a least (smallest) integer; i.e. $\alpha \in A$ such that $\alpha \leq b, \forall b \in A$.**

**OR**

Every non-empty subset of the set of positive integers has a least element.

e.g. $N = \{1, 2, 3, \ldots\} \rightarrow$ smallest element = 1

$A = \{2, 3, 4\} \rightarrow$ smallest element = 2

"Well-ordering principle" is a fundamental tool in Mathematics and is used to:

(i) Prove results using Mathematical Induction

(ii) Prove by contradiction

(iii) Establish the correctness of many algorithms in mathematics and computer science.

**Archimedean Property: If a, b are any two positive integers, then there exists a positive integer n such that $na > b$.**

---

### Divisibility

If a and b are integers with $a \neq 0$, we say that "a divides b" if there is an integer c such that $b = ac$.

Notation: a divides b, we say that 'a is a factor of b'.

The notation $a \mid b$ denotes that a divides b.

If a does not divide b, then we write $a \nmid b$.

e.g. $3 \nmid 7$, since $7/3$ is not an integer.

$3 \mid 12$, since $12/3 = 4$ or $12 = (3)(4)$.

---

### Properties of divisibility

If a, b, c, d, m and n are arbitrary integers, then:

1) $a \mid a$ (Reflexive Property)

**Proof: We know that $a = 1 \cdot a$**

$a = a \cdot 1$, where 1 is an integer

$\Rightarrow a \mid a$

2) If a and b are non-zero integers such that $a \mid b$ and $b \mid a$, then $b = \pm a$ (or) $|b| = |a|$

$a \mid b \Rightarrow b = ak_1$, for some integer $k_1$

$b \mid a \Rightarrow a = bk_2$, for some integer $k_2$

$b = ak_1$

$= (bk_2)k_1$

$b = b(k_2k_1)$

$\Rightarrow b(k_2k_1 - 1) = 0$

Since $k_1, k_2$ are integers and $b \neq 0$, this is possible only if $k_1k_2 = 1$.

or $a = \pm b$ (or) $b = \pm a$ $\Rightarrow |b| = |a|$

3) If a | b and b | c, then a | c. (Transitive property)

If a | b $\Rightarrow b = ak_1$, for some integer $k_1$

If b | c $\Rightarrow c = bk_2$, for some integer $k_2$

$c = bk_2$

$= (ak_1)k_2$

$c = a(k_1k_2)$, where $k_1, k_2$ are integers

$\Rightarrow a \mid c$

4) If a | b and a | c, then

a | mb + nc, where m and n are integers

(Linearity property)

Given a | b $\Rightarrow b = ak_1$, where $k_1$ is some integer.

a | c $\Rightarrow c = ak_2$, where $k_2$ is some integer.

$mb + nc = m(ak_1) + n(ak_2)$

$= a(mk_1 + nk_2)$

Since $mk_1 + nk_2$ is an integer,

$\Rightarrow a \mid mb + nc$

**NOTE:**

If $m = 1$, $n = -1$, then $a \mid b - c$

If $m = 1$, $n = 1$, then $a \mid b + c$

5) If a | b, then a | (bx) for all integers x.

Given a | b $\Rightarrow b = ak_1$, for some integer $k_1$

$bx = (ak_1)x$

$= a(k_1x)$, where $k_1x$ is an integer

$\Rightarrow a \mid bx$

6) If d | n, then d | an where $a \neq 0$. (Multiple Property)

Given d | n $\Rightarrow n = dk_1$, where $k_1$ is an integer

$\Rightarrow an = adk_1$

$= d(ak_1)$

As a, $k_1$ are integers, $ak_1$ is an integer.

$\Rightarrow d \mid an$

7) If ad | an and $a \neq 0$, then d | n. (Cancellation property)

Given ad | an

$\Rightarrow an = (ad)k_1$, where $k_1$ is some integer

$an - adk_1 = 0$

$a(n - dk_1) = 0$

As $a \neq 0$,

$n - dk_1 = 0$

$n = dk_1$

$\Rightarrow d \mid n$

So if $ad \mid an$, then $d \mid n$.

8) $1 \mid n, \ \forall \ n$

As $n = 1 \times n$

$\Rightarrow 1 \mid n$

9) $n \mid 0$ if $n \neq 0$

If $n \neq 0$, then $0 = 0 \times n$

$0 = n \times 0$, where 0 is an integer

$\Rightarrow n \mid 0$

10)

If $a \mid b$ and $a \mid c$, then $a \mid (b + c)$

$b = an$, $c = am$

$b + c = an + am = a(n + m)$

$\Rightarrow a \mid (b + c)$

---

**Problems based on divisibility**

---

**1) If a and b are positive odd integers, prove that $2 \mid (a^2 + b^2)$ but $4 \nmid (a^2 + b^2)$.**

Given a and b are positive odd integers.

$\Rightarrow a = 2m + 1, \quad b = 2n + 1$

$a^2 + b^2 = (2m + 1)^2 + (2n + 1)^2$

$= 4m^2 + 4m + 1 + 4n^2 + 4n + 1$

$= 4(m^2 + n^2 + m + n) + 2$

Clearly it follows that $2 \mid (a^2 + b^2)$, but $4 \nmid (a^2 + b^2)$.

---

**2) If a and b are positive integers with $a \neq 0$ such that $a \mid b$ and $a \mid (b + 2)$, prove that $a = 1$ or $a = 2$.**

Since $a \mid b$ and $a \mid (b + 2)$,

$\Rightarrow a \mid [(b + 2) - b]$ for all integers $x = -1$, $y = 1$

$\Rightarrow a \mid 2$

$\Rightarrow a = 1$ or $a = 2$

---

**3) For any integers a, b, c with $a \neq 0$, prove that if a does not divide bc, then a does not divide b and a does not divide c.**

Suppose $a \mid b$ or $a \mid c$.

Case (i): If $a \mid b$

$\Rightarrow b = ak$, for some integer k

$\Rightarrow bc = a(kc)$

Since k is an integer and c is an integer, kc is an integer.

$\Rightarrow a \mid bc$

Case (ii): Suppose $a \mid c$

$\Rightarrow c = ak_1$, for some integer $k_1$

$\Rightarrow bc = b(ak_1) = a(bk_1)$

$\Rightarrow a \mid bc$

Hence, if $a \mid b$ or $a \mid c$, then $a \mid bc$.

Taking contrapositive of this:

If $a \nmid bc$, then $\neg(a \mid b \text{ or } a \mid c)$

$\neg(a \mid b) \wedge \neg(a \mid c)$

$\Rightarrow$ If $a \nmid bc$, then $a \nmid b$ and $a \nmid c$

---

**4) Show that there do not exist integers x, y, z such that**

$14x - 35y + 42z = 90$

Suppose there exist integers x, y, z such that $14x - 35y + 42z = 90$.

Since 7 divides all of $14x$, $-35y$ and $42z$, it follows that 7 divides $14x - 35y + 42z$. But 7 does not divide 90. This is a contradiction. Hence there are no integers x, y, z such that $14x - 35y + 42z = 90$.

---

**5) If, for some integers a and b, the integer 5a + 8b is a multiple of 12, prove that the integer 19a + 4b is also a multiple of 12.**

$$19a + 4b = 19a + 5a - 5a + 8b - 8b$$
$$= 24a + 12b - (5a + 8b)$$
$$= 12(2a + b) - (5a + 8b)$$

$12(2a + b)$ is a multiple of 12 and $(5a + 8b)$ is also a multiple of 12.

$\Rightarrow 12(2a + b) - (5a + 8b)$ is also a multiple of 12.

---

**6) If a, b, c are integers such that 5a + 7b + 11c is a multiple of 31, prove that 21a + 17b + 4c is also a multiple of 31.**

$$21a + 17b + 4c = 31(a + b + c) - 2(5a + 7b + 11c)$$

$31(a + b + c)$ is divisible by 31 and $2(5a + 7b + 11c)$ is also divisible by 31.

$\Rightarrow 21a + 17b + 4c$ is divisible by 31.

---

**7) Prove that $3 \mid (7^n - 4^n)$ for every positive integer n.**

For $n = 1$,

$3 \mid (7^1 - 4^1) = 3 \mid 3$

$\Rightarrow$ The result is true for $n = 1$.

Assume that the result is true for $n = k$.

$3 \mid (7^k - 4^k)$

Consider,

$7^{k+1} - 4^{k+1} = 7 \cdot 7^k - 4 \cdot 4^k$

$= 3 \cdot 7^k + 4 \cdot 7^k - 4 \cdot 4^k$

$= 3 \cdot 7^k + 4(7^k - 4^k)$

$3 \mid 3 \cdot 7^k$ and $3 \mid 4(7^k - 4^k)$

$\Rightarrow 3 \mid [7^{k+1} - 4^{k+1}]$

---

## Prime numbers

An integer $p > 1$ whose only divisors are 1 and p is called a prime number.

An integer > 1 which is not a prime number is called a composite number.

e.g. Prime numbers from 1 to 100 are

2, 3, 5, 7
11, 13, 17, 19
23, 29
31, 37
41, 43, 47
53, 59
61, 67
71, 73, 79
83, 89
97

e.g. The integers 4, 6, 8, 9, 10, 12, 14, 15 are composite numbers.

If n is a composite number, then n has divisors other than $\pm 1$ and $\pm n$.

$\Rightarrow n = n_1 n_2$ for some integers $n_1$, $n_2$ with $0 < n_1 < n$ and $0 < n_2 < n$

## Division Algorithm

Let a and b be integers with $b \neq 0$. Then there exist unique integers q and r such that

$a = bq + r$, with $0 \leq r < |b|$

## Common divisor

Given two integers 'a' and 'b', if there is an integer 'd' such that $d \mid a$ and $d \mid b$, then d is called a common divisor of 'a' and 'b'.

## Greatest common divisor

Suppose 'a' and 'b' are integers. If there is an integer $d > 0$ such that

(i) $d \mid a$ and $d \mid b$

(ii) every common divisor of 'a' and 'b' divides d,

then we say 'd' is the greatest common divisor (gcd) of a and b. It is denoted by gcd(a, b) or (a, b).

e.g. 1) GCD of 24 and 36

Divisors of 24 are 1, 2, 3, 4, 6, 8, 12, 24.

Divisors of 36 are 1, 2, 3, 4, 6, 9, 12, 18, 36.

Common divisors of 24 and 36 are

1, 2, 3, 4, 6, 12

gcd(24, 36) = 12

2) What is the gcd of 17 and 22?

Divisors of 17 are 1, 17.

Divisors of 22 are 1, 2, 11, 22.

Common divisors are 1.

gcd = 1

3) Find gcd(12, 18).

Divisors of 12 are 1, 2, 3, 4, 6, 12.

Divisors of 18 are 1, 2, 3, 6, 9, 18.

C.D. are 1, 2, 3, 6.

GCD is 6

## Co-prime numbers (or) Relatively prime numbers

If gcd(a, b) = 1, then we say that a and b are relatively prime integers.

e.g. 12, 35 are relatively prime.

As divisors of 12 are 1, 2, 3, 4, 6, 12

Divisors of 35 are 1, 5, 7.

C.D. = 1

G.C.D. = 1

**Examples of relatively prime integers**

(i) a = 12, b = 35

$35 = 12(2) + 11$

$12 = 11(1) + 1$

$11 = 1(11) + 0$

gcd(12, 35) = 1

⇒ 12, 35 are relatively prime.

(ii) a = 49, b = 18

$49 = 18(2) + 13$

$18 = 13(1) + 5$

$13 = 5(2) + 3$

$5 = 3(1) + 2$

$3 = 2(1) + 1$

$2 = 1(2) + 0$

gcd(49, 18) = 1

⇒ 49, 18 are co-primes.

(iii) a = 21, b = 64

$64 = 21(3) + 1$

$21 = 1(21) + 0$

gcd(21, 64) = 1

⇒ 21, 64 are relatively prime.

(iv) a = 28, b = 45

$45 = 28(1) + 17$

$28 = 17(1) + 11$

$17 = 11(1) + 6$

$11 = 6(1) + 5$

$6 = 5(1) + 1$

$5 = 1(5) + 0$

$\gcd(28, 45) = 1$

$\Rightarrow$ 28, 45 are relatively prime.

NOTE: For positive integers a and b, the gcd is the smallest positive integer which can be written in the form $ax + by$ for some integers x and y.

## Properties of co-prime (or) relatively prime numbers

1) Any two distinct primes are relatively prime. e.g. $(7, 13) = 1$
2) Any two consecutive integers are relatively prime. e.g. $(9, 10) = 1$
3) The integers a and b are relatively prime if and only if 1 can be expressed as a linear combination of a and b.
4) Suppose a and b are relatively prime and $a \mid bc$, then $a \mid c$.
5) Suppose a, b and c are integers such that a and b are relatively prime and $a \mid c$ and $b \mid c$. Then $ab \mid c$.
6) Suppose a prime p divides a product ab. Then $p \mid a$ or $p \mid b$.

**Problem**

Show that, for any positive integer m, the numbers $3m + 2$ and $11m + 7$ are relatively prime.

**Sol:**

$$1 = 11(3m + 2) - 3(11m + 7)$$

This shows that 1 is a linear combination of $3m + 2$ and $11m + 7$.

Therefore, $3m + 2$ and $11m + 7$ are relatively prime.

NOTE: For positive integers a and b, the gcd is the smallest positive integer which can be written in the form $ax + by$ for some integers x and y.

e.g. $a = 12$, $b = 18$, $\gcd(12, 18) = 6$.

## Properties of gcd

The gcd has the following properties:

(1) $\gcd(a, b) = \gcd(b, a)$ (Commutative law)

**Sol: Let $\gcd(a, b) = d$.**

d is the smallest positive integer which can be written in the form

$d = ax + by$, for some integers x and y

$d = by + ax$

d is the smallest positive integer such that $d = by + ax$, for some integers x and y.

$\Rightarrow \gcd(b, a) = d$

$\Rightarrow \gcd(a, b) = \gcd(b, a)$

(2) $\gcd(a, (b, c)) = \gcd((a, b), c)$ (Associative law)

(or) $(a, (b, c)) = ((a, b), c)$

**Proof: Let $d = \gcd(a, \gcd(b, c))$. Then $d \mid a$ and $d \mid \gcd(b, c)$, so $d \mid a$, $d \mid b$ and $d \mid c$.**

**Similarly, let $e = \gcd(\gcd(a, b), c)$. Then $e \mid a$, $e \mid b$ and $e \mid c$.**

**Thus d and e are both the greatest positive common divisor of a, b and c. Therefore $d = e$.**

**Hence $\gcd(a, \gcd(b, c)) = \gcd(\gcd(a, b), c)$.**

(3) $(ac, bc) = |c|(a, b)$ (Distributive law)

Let $d = (a, b)$.

$\Rightarrow d \mid a$ and $d \mid b$

$a = dm$, $b = dn$, for some integers m, n

$ac = dcm$ and $bc = dcn$

Hence $|c|d$ divides both ac and bc.

$(ac, bc) = |c|(a, b)$

**Proof for $c > 0$**

Let $g = (a, b)$ and $G = (ac, bc)$. We want to prove that $G = cg$.

1) Show that $cg \leq G$

By definition of gcd, $g \mid a$ and $g \mid b$.

Multiplying by c, $cg \mid ac$ and $cg \mid bc$.

As cg is a common divisor of ac and bc, it must be less than or equal to the greatest common divisor of ac and bc.

$\Rightarrow cg \leq G$ ...(1)

2) To show that $G \leq cg$

As $g = (a, b)$, there exist integers x and y such that

$g = ax + by$

$cg = acx + bcy$

We know that $G = (ac, bc)$.

$\Rightarrow G \mid ac$ and $G \mid bc$

$\Rightarrow G \mid (acx + bcy)$

$\Rightarrow G \mid cg$

$\Rightarrow G \leq cg$ ...(2)

From (1) and (2), $G = cg$.

**Proof for $c < 0$**

If c is negative, let $k = -c = |c|$, where $k > 0$.

Using the property $(x, y) = (|x|, |y|)$,

$(ac, bc) = (|ac|, |bc|)$

$= (|a|k, |b|k)$

Applying the result for positive integers proved above:

$(|a|k, |b|k) = k(|a|, |b|) = |c|(a, b)$

So, for any non-zero integer c,

$(ac, bc) = |c|(a, b)$

(4) $(a, 1) = (1, a) = 1$ and $(a, 0) = (0, a) = |a|$

(i) Let $(a, 1) = d$.

$\Rightarrow d \mid a$ and $d \mid 1$

As $d \mid 1$, $d = 1$.

$\Rightarrow (a, 1) = 1$ and $(1, a) = 1$

(ii) Let $(a, 0) = d$.

$\Rightarrow d = ax + 0y$, for some integers x, y

$d = ax$

$\Rightarrow a \mid d$ ...(1)

As $(a, 0) = d$,

$\Rightarrow d \mid a$ ...(2)

From (1) and (2), $|d| = |a|$.

$\Rightarrow (a, 0) = (0, a) = |a|$

---

## Euclidean Algorithm

---

Statement: Given two positive integers a and b with a > b, an algorithm known as Euclid's algorithm may be employed to obtain the gcd of a and b.

According to this algorithm, gcd(a, b) is obtained by applying the division algorithm repeatedly to a and b and obtaining a set of remainders as shown below.

Let $a = r_0$, $b = r_1$

As $a > b$, $r_0 > r_1$

$r_0 = r_1q_1 + r_2, \quad 0 < r_2 < r_1$

$r_1 = r_2q_2 + r_3, \quad 0 < r_3 < r_2$

$r_2 = r_3q_3 + r_4, \quad 0 < r_4 < r_3$

$\vdots$

$r_{n-2} = r_{n-1}q_{n-1} + r_n, \quad 0 < r_n < r_{n-1}$

$r_{n-1} = r_nq_n + 0$

By division algorithm: dividend = divisor × quotient + remainder.

We stop the process when the remainder becomes zero. The last non-zero remainder $r_n$ is the gcd(a, b).

---

## Problems on gcd

---

**1) Find the gcd of a = 258 and b = 24. Also express the gcd as a linear combination of a and b.**

$258 = 24(10) + 18$

$24 = 18(1) + 6$

$18 = 6(3) + 0$

Since the remainder is zero, we stop the process.

The last non-zero remainder is 6.

$\Rightarrow \gcd(258, 24) = 6$

Now, by working backwards, we find that

$$\begin{aligned} \gcd = 6 &= 24 - 18(1) \\ &= 24 - [1 \times 258 - 24 \times 10] \\ &= 24 - 258 + 24 \times 10 \end{aligned}$$

$= (-1) \times 258 + 11 \times 24$

Thus gcd(258, 24) = 6 is expressed as a linear combination of 258 and 24.

---

**2) Find the gcd of 250 and 111. Express the gcd as a linear combination of these numbers.**

250 = 111(2) + 28

111 = 28(3) + 27

28 = 27(1) + 1

27 = 1(27) + 0

gcd(250, 111) = 1

$1 = 28 - 1 \times 27$

$= 28 - [1 \times \{111 - (28 \times 3)\}]$

$= 28 - 111 + 28 \times 3$

$= 28 \times 4 - 111$

$= [250 - (111 \times 2)] \times 4 - 111$

$= 250 \times 4 - 8 \times 111 - 111$

$= 250 \times 4 - 9 \times 111$

gcd(250, 111) is expressed as a linear combination of 250 and 111.

---

**3) Find the gcd of 826 and 1890 using Euclidean algorithm and express the gcd as a linear combination of 826 and 1890.**

1890 = 826(2) + 238

826 = 238(3) + 112

238 = 112(2) + 14

112 = 14(8) + 0

gcd = 14

$14 = 238 - (112 \times 2)$

$= 238 - [2 \times \{826 - 3 \times 238\}]$

$= 238 - 2 \times 826 + 6 \times 238$

$= 238 \times 7 - 2 \times 826$

$= [1890 - 826 \times 2] \times 7 - 2 \times 826$

$= 1890 \times 7 - 14 \times 826 - 2 \times 826$

$= 1890 \times 7 - 16 \times 826$

---

**4) By using Euclidean algorithm find the gcd of 1769 and 2378 and also find x and y such that**

(1769, 2378) = 1769x + 2378y

2378 = 1769(1) + 609

1769 = 609(2) + 551

609 = 551(1) + 58

551 = 58(9) + 29

58 = 29(2) + 0

gcd = 29

$29 = 551 - 58 \times 9$

$= 551 - [609 - 551 \times 1] \times 9$

$= 551 - 609 \times 9 + 551 \times 9$

$= 551 \times 10 - 609 \times 9$

$= [1769 - (609 \times 2)] \times 10 - 609 \times 9$

$= 1769 \times 10 - 609 \times 20 - 609 \times 9$

$= 1769 \times 10 - 609 \times 29$

$= 1769 \times 10 - [2378 - 1769 \times 1] \times 29$

$= 1769 \times 10 - 2378 \times 29 + 1769 \times 29$

$= 1769 \times 39 - 2378 \times 29$

$= 1769x + 2378y$

$\Rightarrow x = 39, y = -29$

---

## Problems for Practice

---

**1) Find the gcd of 1479 and 272 using Euclidean algorithm and express the gcd as a linear combination of 1479 and 272.**

gcd = 17

$17 = 1479x + 272y$

---

**2) 12378, 3054**

gcd = 6

$6 = 132 \times 12378 - 535 \times 3054$

---

**3) gcd(56, 72) = 56x + 72y, find x, y**

gcd(56, 72) = 8

$8 = 56x + 72y$

$x = 4, y = -3$

---

**4) gcd(24, 138) = 24x + 138y, find x, y**

gcd(24, 138) = 6

$6 = 24 \times 6 + 138(-1)$

$x = 6, y = -1$

---

**5) gcd(23, 37) = 23a + 37b, find a, b**

gcd(23, 37) = 1

$a = -8, b = 5$

---

## Fundamental Theorem of Arithmetic

---

Every integer $n > 1$ can be represented as a product of prime factors in only one way, apart from the order of factors.

**Proof: We prove the theorem by induction on n.**

For $n = 2$, the theorem is true.

2 itself is a prime number

Assume that the theorem is true for $m > 1$ and less than n.
If n is a prime, then there is nothing to prove.
If n is composite, then there is a positive divisor d with $d \neq 1$ and $d \neq n$; thus $n = cd$, where $c \neq 1$, $c \neq n$.
Since $c < n$ and $d < n$, by hypothesis c and d can be written as products of prime numbers.

$\Rightarrow$ n is a product of prime numbers.

To prove uniqueness, we use Euclid's Lemma: if a prime p divides a product ab, then $p \mid a$ or $p \mid b$.
Proof of the lemma: If $p \nmid a$, then $\gcd(p, a) = 1$. Hence there exist integers x and y such that $px + ay = 1$.
Multiplying by b gives $pbx + aby = b$. Since p divides both terms on the left, $p \mid b$. Thus $p \mid a$ or $p \mid b$.
Now suppose that n has two prime-factor representations:

$n = p_1 p_2 \ldots p_k$

$n = q_1 q_2 \ldots q_l$

where $p_i$'s and $q_j$'s are primes.

$p_1 \mid n \Rightarrow p_1 \mid (q_1 q_2 \ldots q_l)$

$\Rightarrow p_1$ divides one of the $q_j$'s, say $q_i$

By the theorem: If p is a prime and $p \mid ab$, then $p \mid a$ or $p \mid b$.
If $p_1 \mid q_i$, then $p_1 = q_i$.
[Since p and q are primes and $p \mid q$, then $p = q$.]
Proceeding like this, we find that every p must equal some q and conversely q must be equal to some p.
This proves the required uniqueness of the result.
NOTE: The above theorem is also known as the Unique Factorization Theorem.
The representation of $n > 1$ as a product of primes is referred to as the canonical representation of n.

**Examples**

1) Express 10! as product of primes.

$10! = 10 \times 9 \times 8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1$
$= 2 \times 5 \times 3^2 \times 2^3 \times 7 \times 2 \times 3 \times 5 \times 2^2 \times 3 \times 2$
$= 2^8 \times 3^4 \times 5^2 \times 7$

2) Express 3105 as product of primes.

$3105 = 3^3 \times 5 \times 23$

3) Find the prime factorization of

(i) 100 (ii) 641 (iii) 999 (iv) 1024 (v) 7007

$100 = 2^2 \times 5^2$

$641 = 641$

$999 = 3^3 \times 37$

$1024 = 2^{10}$

$7007 = 7 \times 7 \times 11 \times 13$

**Problems**

Using F.T.A., express the following numbers as a product of prime numbers:

1) 5544

$5544 = 2^3 \times 3^2 \times 7 \times 11$

2) 4849845

$4849845 = 3 \times 5 \times 7 \times 11 \times 13 \times 17 \times 19$

3) 9999

$9999 = 3^2 \times 11 \times 101$

---

## Euclid's Lemma

If $a \mid bc$ and if $(a, b) = 1$, then $a \mid c$.

**Proof: Since (a, b) = 1, we can write**

$1 = ax + by$, where x, y are integers

$c = acx + bcy$

$a \mid ac$ and it is given that $a \mid bc$.

$\Rightarrow a \mid acx$ and $a \mid bcy$

$\Rightarrow a \mid acx + bcy$

$\Rightarrow a \mid c$

---

## Prime number

An integer $n > 1$ is called a prime number if the only positive divisors are 1 and n. If $n > 1$ and if n is not prime, then n is called a composite number.

## Euclid's theorem for Prime Numbers

There are infinitely many prime numbers.

**Proof: (By contradiction)**

Assume that there are finitely many prime numbers. Then we can say that there are n prime numbers and we can write them down, in order:

$2 = p_1 < p_2 < p_3 < \dots < p_n$

Let the list of all prime numbers be $p_1, p_2, p_3, \dots, p_n$.

Let $N = 1 + p_1p_2p_3 \dots p_n$

Since $N > p_n$ and $p_n$ is the largest prime number, N is not prime.

By the Fundamental Theorem of Arithmetic,

N can be written as a product of primes, i.e., N must have a prime factor.

This means one of the primes in our list must divide N.

$\Rightarrow$ There exists an integer i with $1 \le i \le n$ such that $p_i$ divides N.

Since $p_i \mid N$ and $p_i \mid p_1p_2p_3 \dots p_i \dots p_n$,

$\Rightarrow p_i \mid [N - (p_1p_2p_3 \dots p_i \dots p_n)]$

$\Rightarrow p_i \mid 1$

But $p_i \ge 2$; it is impossible that $p_i$ divides 1.

So we get a contradiction.

Hence our assumption that there are finitely many primes is false.

$\Rightarrow$ There are infinitely many primes.

---

## Fermat numbers

The numbers which are of the form $2^{2n} + 1$, $n \geq 0$, are called Fermat numbers.

$F_n = 2^{2n} + 1$, $n \geq 0$

$F_0 = 2^1 + 1 = 3$

$F_1 = 2^2 + 1 = 5$

$F_2 = 2^4 + 1 = 17$

$F_3 = 2^8 + 1 = 257$

$F_4 = 2^{16} + 1 = 65537$

$F_5 = 2^{32} + 1 = 4294967297$

NOTE: The Fermat numbers $F_0$, $F_1$, $F_2$, $F_3$ and $F_4$ are prime, whereas $F_5$ is composite.

These numbers are also used in plane geometry.

Gauss proved that if $F_n$ is a prime, say $F_n = p$, then a regular polygon of p sides can be constructed with straight edge and compass.

No Fermat prime beyond $F_4$ is currently known. In fact, for $5 \leq n \leq 16$, each Fermat number $F_n$ is composite.

---

**Problem 1**

Use Fermat factorization method to factorize n = 119143.

We know that

$(345)^2 = 119025 < 119143 < (346)^2 = 119716$

$(346)^2 - 119143 = 573$

$(347)^2 - 119143 = 1266$

$(348)^2 - 119143 = 1961$

$(349)^2 - 119143 = 2658$

$(350)^2 - 119143 = 3357$

$(351)^2 - 119143 = 4058$

$(352)^2 - 119143 = 4761 = 69^2$

$(352)^2 - (69)^2 = 119143$

$\Rightarrow 119143 = (352 + 69)(352 - 69)$

$= 421 \times 283$

This is the required factorization.

---

**Problem 2**

Use Fermat factorization method to factorize n = 23449.

As $\sqrt{23449} \approx 153.13$, so $153 < \sqrt{23449} < 154$.

$(153)^2 < 23449 < (154)^2$

$(154)^2 - 23449 = 23716 - 23449 = 267$

$(155)^2 - 23449 = 576 = 24^2$

$\Rightarrow (155)^2 - (24)^2 = 23449$

$23449 = (155 + 24)(155 - 24)$

$= 179 \times 131$

This is the required factorization.

## Euclidean Algorithm

**Theorem: Statement**

Given two positive integers a and b, where a > b.

Let $a = r_0$, $b = r_1$, and apply the division algorithm repeatedly to obtain a set of remainders $r_2, r_3, ..., r_n$ after successive divisions by the relations

$r_0 = r_1q_1 + r_2, \quad 0 < r_2 < r_1$

$r_1 = r_2q_2 + r_3, \quad 0 < r_3 < r_2$

$r_2 = r_3q_3 + r_4, \quad 0 < r_4 < r_3$

$\vdots$

$r_{n-2} = r_{n-1}q_{n-1} + r_n, \quad 0 < r_n < r_{n-1}$

$r_{n-1} = r_nq_n + r_{n+1}$, where $r_{n+1} = 0$

Then $r_n$, the last non-zero remainder in this process, is (a, b), the gcd of a and b.

**Proof: Part 1**

Given a and b are positive integers, where $a = r_0$, $b = r_1$.

By the definition of $r_i$'s (i = 1, 2, 3, ...), we observe that $r_i$ is a non-negative and decreasing sequence. After a finite number of steps,

$r_{n-1} = r_nq_n + r_{n+1}$, where $r_{n+1} = 0$

$r_{n-1} = r_nq_n$

$\Rightarrow r_n \mid r_{n-1}$

From the previous equation,

$r_{n-2} = r_{n-1}q_{n-1} + r_n$

$r_{n-2} = r_nq_nq_{n-1} + r_n$

$r_{n-2} = (q_nq_{n-1} + 1)r_n$

$\Rightarrow r_n \mid r_{n-2}$

Similarly, from other relations,

$r_n \mid r_{n-3}, r_n \mid r_{n-4}, ..., r_n \mid r_1, r_n \mid r_0$

As $a = r_0$ and $b = r_1$,

$\Rightarrow r_n \mid b, r_n \mid a$

$\Rightarrow r_n$ is a common divisor of a and b.

**Part 2**

Let p be any common divisor of a and b.

$\Rightarrow p \mid a$

$\Rightarrow p \mid b$

We have to prove that $p \mid r_n$.

Since $r_0 = r_1q_1 + r_2$

$r_2 = r_0 - r_1q_1$

As $p \mid r_0$ and $p \mid r_1$,

$\Rightarrow p \mid r_2$

Similarly, $p \mid r_3, \ldots, p \mid r_{n-1}, p \mid r_n$.

Every common divisor of a and b divides $r_n$.

$\Rightarrow \gcd(a, b) = r_n$

## CONGRUENCES

The word congruence was introduced by Karl F. Gauss. It is denoted by "≡".

If m is a positive integer, then two numbers a, b are said to be congruent modulo m if $m \mid (a - b)$.

$a \equiv b \pmod{m}$ if and only if $a - b = mq$, for some $q \in Z$

**Examples:**

1) $3 \equiv 24 \pmod{7}$, since $7 \mid (3 - 24) \Rightarrow 7 \mid -21$
2) $-31 \equiv 11 \pmod{7}$, since $7 \mid (-31 - 11) \Rightarrow 7 \mid -42$
3) $-15 \equiv -64 \pmod{7}$, since $7 \mid (-15 + 64) \Rightarrow 7 \mid 49$
4) $6 \equiv 1 \pmod{5}$, since $5 \mid (6 - 1) \Rightarrow 5 \mid 5$

**NOTE:**

1) Any two numbers are congruent modulo 1.

2) Any two integers are congruent modulo 2 if both are even or both are odd. Also, every integer is congruent to its negative modulo 2, since $a - (-a) = 2a$.

3) If a is any integer, then by the division algorithm,

$a = mq + r$, where $0 \le r < m$, $m > 0$

$\Rightarrow a - r = mq$

$\Rightarrow m \mid (a - r)$

$\Rightarrow a \equiv r \pmod{m}$, where $r = 0, 1, 2, \ldots, (m - 1)$

### Linear congruence

A linear congruence is an equation of the form $ax \equiv b \pmod{m}$, where a, b are integers, m is a positive integer ($m > 0$), and x is an unknown integer to be found.

### Solution of linear congruence

The linear congruence $ax \equiv b \pmod{m}$ has a solution if $\gcd(a, m)$ divides b.

NOTE: If $\gcd(a, m) \nmid b$, then it has no solution.

The congruence means

$m \mid (ax - b)$

$\Rightarrow ax - b = mk$, for some integer k

## Number of solutions

1) If gcd(a, m) = d and d | b, then there are exactly d incongruent solutions modulo m.

Two solutions $x_1$ and $x_2$ to a linear congruence are considered incongruent modulo m if they leave different remainders when divided by m.

In simpler terms, they are completely distinct when compared within the standard set of remainders {0, 1, 2, ⋯, m − 1}.

Congruent solutions: $x_1 \equiv x_2 \pmod{m}$. This means $x_1 - x_2$ is exactly divisible by m. They represent the same family of numbers.

Incongruent solutions: $x_1 \not\equiv x_2 \pmod{m}$. This means $x_1 - x_2$ is not divisible by m. They represent truly unique answers.

Example: $2x \equiv 4 \pmod{6}$

Find the number of solutions.

a = 2, b = 4, m = 6

gcd(a, m) = gcd(2, 6) = 2

Since 2 | 4, there are exactly 2 incongruent solutions modulo 6.

Test values from 0 to 5, as it is modulo 6.

If x = 2, $2x \equiv 4 \pmod{6}$

$4 \equiv 4 \pmod{6} \Rightarrow 6 \mid (4 - 4) \Rightarrow$ True

If x = 5, $2(5) \equiv 4 \pmod{6}$

$10 \equiv 4 \pmod{6} \Rightarrow 6 \mid (10 - 4) \Rightarrow 6 \mid 6 \Rightarrow$ True

Therefore, the two solutions 2 and 5 are incongruent, since $2 \not\equiv 5 \pmod{6}$, as $6 \nmid (2 - 5)$.

If we continue testing, we find that x = 8 is also a solution as

$2(8) \equiv 4 \pmod{6}$

But $8 \equiv 2 \pmod{6}$. Therefore x = 8 is congruent to x = 2. It is not counted as a new, unique solution.

2) If $d \nmid b$, there is no solution.

Example:

$6x \equiv 7 \pmod{14}$

a = 6, b = 7, m = 14

gcd(6, 14) = 2

Since $2 \nmid 7$, the congruence has no solution.

3) The solutions can be obtained by

$x = x_0 + (m/d)t$, where $x_0$ is an initial solution and t = 0, 1, 2, ⋯, (d − 1)

## Linear Diophantine equation

If $ax \equiv b \pmod{m}$ is a linear congruence, it can be expressed as $ax - my = b$. This is known as a Diophantine equation.

$ax \equiv b \pmod{m}$

$\Rightarrow m \mid (ax - b)$

$\Rightarrow ax - b = my$

$\Rightarrow ax - my = b$

Example: Find all possible values of the linear congruence

$$9x \equiv 12 \pmod{15} \quad \ldots(1)$$

Comparing with $ax \equiv b \pmod{m}$:

$a = 9, b = 12, m = 15$

$\gcd(a, m) = \gcd(9, 15) = 3$

$3 \mid 12$

So equation (1) has 3 solutions.

Solution set:

$$x = x_0 + (m/d)t, \quad t = 0, 1, 2, \cdots, (d - 1)$$

We know that the linear Diophantine equation of (1) is

$$9x - 15y = 12 \quad \ldots(2)$$

$\gcd(9, 15) = 3$

$15 = 9(1) + 6$

$9 = 6(1) + 3$

$6 = 3(2) + 0$

$$\begin{aligned} 3 &= 9 - 6(1) \\ &= 9 - [15 - 9] \\ &= 2\cdot 9 - 15\cdot 1 \end{aligned}$$

Multiplying by 4 on both sides:

$$12 = 9(8) - 15(4)$$

Comparing with $9x - 15y = 12$,

$x_0 = 8, y_0 = 4$

$x = 8 + (15/3)t, \quad t = 0, 1, 2$

$x = 8 + 5t$

If $t = 0, x = 8$

If $t = 1, x = 8 + 5 = 13$

If $t = 2, x = 8 + 10 = 18 \equiv 3 \pmod{15}$

Therefore, the least non-negative incongruent solutions are $x \equiv 3, 8, 13 \pmod{15}$.

Method 2:

Given that

$$9x \equiv 12 \pmod{15} \quad \ldots(1)$$

Comparing $ax \equiv b \pmod{m}$:

$a = 9, b = 12, m = 15$

$\gcd(a, m) = \gcd(9, 15) = 3$

$15 = 9(1) + 6$

$9 = 6(1) + 3$

$6 = 3(2) + 0$

$$\begin{aligned} 3 &= 9 - 6(1) \\ &= 9 - [15 - 9] \end{aligned}$$

$= 2 \cdot 9 - 15$

$\Rightarrow 3 \mid 12$

The given congruence $9x \equiv 12 \pmod{15}$ will have 3 distinct solutions.

$x = x_0 + (m/d)t$, where $t = 0, 1, 2$

$x = x_0 + 5t$

Testing values:

For $x = 1$: 15 does not divide $9 - 12 = -3$

For $x = 2$: 15 does not divide $18 - 12 = 6$

For $x = 3$: 15 divides $27 - 12 = 15$

So let $x_0 = 3$ be the initial solution of the given congruence.

$x = x_0 + (m/d)t, \ t = 0, 1, 2$

$x = 3 + (15/3)t$

$x = 3 + 5t$

$x = 3$ when $t = 0$

$x = 8$ when $t = 1$

$x = 13$ when $t = 2$

Hence $x \equiv 3, 8$ and $13 \pmod{15}$ are the three incongruent solutions.

2) Find all possible sets of the linear congruence

$7x \equiv 1 \pmod{31} \quad \ldots(1)$

$a = 7, b = 1, m = 31$

$\gcd(a, m) = \gcd(7, 31) = 1$

Therefore equation (1) will have only one solution.

$31 = 7(4) + 3$

$7 = 3(2) + 1$

$3 = 1(3) + 0$

$1 = 7 - 3(2)$

$= 7 - [31 - 7(4)]2$

$= 7 - 2 \cdot 31 + 8 \cdot 7$

$= 9 \cdot 7 - 2 \cdot 31$

We know that the Diophantine equation of (1) is

$7x - 31y = 1$

Comparing, we get

$x_0 = 9, y_0 = 2$

$x = x_0 + (m/d)t$, where $t = 0$

$x = 9$

3) Find all solutions of

$14x \equiv 12 \pmod{18} \quad \ldots(1)$

Comparing with $ax \equiv b \pmod{m}$:

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

$\gcd(a, m) = \gcd(14, 18) = 2 = d$

$18 = 14(1) + 4$

$14 = 4(3) + 2$

$4 = 2(2) + 0$

As $2 \mid 12$, equation (1) will have 2 distinct solutions.

$$
\begin{aligned}
2 &= 14 - 4(3) \\
&= 14 - [18 - 14]3 \\
&= 14 - 18\cdot 3 + 14\cdot 3 \\
&= 14\cdot 4 - 18\cdot 3 \quad \ldots(2)
\end{aligned}
$$

The Diophantine equation of (1) is

$$14x - 18y = 12$$

Multiplying both sides of equation (2) by 6:

$$12 = 14(24) - 18(18)$$

$x_0 = 24$, $y_0 = 18$

Solution set:

$x = x_0 + (m/d)t, \; t = 0, 1$

$x = 24 + (18/2)t$

$x = 24 + 9t$

For $t = 0$, $x = 24 \equiv 6 \pmod{18}$

For $t = 1$, $x = 33 \equiv 15 \pmod{18}$

Therefore, the two incongruent solutions are $x \equiv 6$ and $x \equiv 15 \pmod{18}$.

Characterize whether the following linear congruences have solutions or not. If so, find all possible solutions.

4)

$$9x \equiv 6 \pmod{15} \quad \ldots(1)$$

$a = 9$, $b = 6$, $m = 15$

$\gcd(a, m) = \gcd(9, 15) = 3$

$3 \mid 6$

Therefore equation (1) will have 3 distinct solutions.

$15 = 9(1) + 6$

$9 = 6(1) + 3$

$6 = 3(2) + 0$

$$
\begin{aligned}
3 &= 9 - 6 \\
&= 9 - [15 - 9] \\
&= 2\cdot 9 - 15 \quad \ldots(2)
\end{aligned}
$$

The Diophantine equation of (1) is

$$9x - 15y = 6$$

Comparing equation (2) with the above form and multiplying by 2:

$$6 = 9(4) - 15(2)$$

$x_0 = 4$, $y_0 = 2$

$x = x_0 + (m/d)t$, where $t = 0, 1, 2$

$x = 4 + 5t$

$x = 4 + 5t, \; t = 0, 1, 2$

$x = 4$

$x = 9$

$x = 14$

5)

$$10x \equiv 15 \pmod{45} \quad \ldots(1)$$

Comparing with $ax \equiv b \pmod{m}$:

$a = 10, b = 15, m = 45$

$\gcd(a, m) = \gcd(10, 45) = 5 = d$

As $5 \mid 15$, equation (1) has 5 distinct solutions.

$45 = 10(4) + 5$

$10 = 5(2) + 0$

$5 = 45 - 10(4)$

$\quad = 45\cdot1 - 10\cdot4 \quad \ldots(2)$

The Diophantine equation of (1) is

$$10x - 45y = 15$$

Using equation (2):

$10(-4) - 45(-1) = 5$

$10(-4 \times 3) - 45(-1 \times 3) = 15$

$10(-12) - 45(-3) = 15$

$x_0 = -12,\ y_0 = -3$

$x = x_0 + (m/d)t, \ \ t = 0, 1, 2, 3, 4$

$x = -12 + 9t$

$x = -12, -3, 6, 15, 24$; equivalently, $x \equiv 6, 15, 24, 33, 42 \pmod{45}$.

6)

$$183x \equiv 15 \pmod{31} \quad \ldots(1)$$

$a = 183, b = 15, m = 31$

$\gcd(a, m) = \gcd(183, 31) = 1 = d$

$1 \mid 15$

Therefore equation (1) has only one distinct solution.

$183 = 31(5) + 28$

$31 = 28(1) + 3$

$28 = 3(9) + 1$

$3 = 1(3) + 0$

$1 = 28 - 3(9)$

$\quad = 28 - [31 - 28]9$

$\quad = 28 - 31\cdot9 + 28\cdot9$

$\quad = 28\cdot10 - 31\cdot9$

$\quad = [183 - 31\cdot5]10 - 31\cdot9$

$\quad = 183\cdot10 - 31\cdot50 - 31\cdot9$

$\quad = 183\cdot10 - 31\cdot59 \quad \ldots(2)$

The Diophantine equation of (1) is

$$183x - 31y = 15$$

Multiplying equation (2) by 15:

$$15 = 183(150) - 31(885)$$

$x_0 = 150$

**Solution continued**

$x = x_0 + (m/d)t$

$= 150 + (31/1)t$, where $t = 0,1,2,\cdots,(d-1)$

$t = 0$

$x = 150$

$x \equiv 150 \equiv 26 \pmod{31}$

---

**System of linear congruences**

---

A linear congruence which is of the form $ax + by \equiv r \pmod{m}$, $cx + dy \equiv s \pmod{m}$ is known as a system of linear congruences in two variables, where $m$ is a positive integer and $a, b, c, d, r, s \in \mathbb{Z}$.

The necessary and sufficient condition for the system of linear congruences to have a unique solution is

$$\gcd(ad - bc, m) = 1$$

1) Find the solution of the system of linear congruences

$3x + 4y \equiv 5 \pmod{13}$ ...(1)

$2x + 5y \equiv 7 \pmod{13}$ ...(2)

Comparing equation (1) with $ax + by \equiv r \pmod{m}$:

$a = 3, b = 4, r = 5, m = 13$

Comparing equation (2) with $cx + dy \equiv s \pmod{m}$:

$c = 2, d = 5, s = 7, m = 13$

$ad - bc = 3(5) - 4(2) = 15 - 8 = 7$

$\gcd(ad - bc, m) = \gcd(7,13) = 1$

Therefore the given system has a unique solution.

Multiply equation (1) by 5 and equation (2) by 4:

$15x + 20y \equiv 25 \pmod{13}$

$8x + 20y \equiv 28 \pmod{13}$

Eliminating y:

$7x \equiv -3 \pmod{13}$

This is a linear congruence in one variable with $a = 7$, $b = -3$, $m = 13$.

$\gcd(7,13) = 1$

So equation has one solution.

From the Diophantine equation

$7x - 13y = -3$ ...(4)

$13 = 7(1) + 6$

$7 = 6(1) + 1$

$6 = 1(6) + 0$

$1 = 7 - 6$

$= 7 - (13 - 7)$

$= 2\cdot 7 - 13$

Multiplying by $-3$:

$-3 = 7(-6) - 13(-3) \quad \ldots(5)$

$x_0 = -6$, $y_0 = -3$

Let the initial value be $x_0 = -6$.

$x = x_0 + (m/d)t$, where $t = 0$

$x = -6$

Substitute $x = -6$ in equation (1).

Eliminating x:

Multiply equation (1) by 2 and equation (2) by 3:

$6x + 8y \equiv 10 \pmod{13}$

$6x + 15y \equiv 21 \pmod{13}$

$-7y \equiv -11 \pmod{13}$

$7y \equiv 11 \pmod{13} \quad \ldots(6)$

$\gcd(7,13) = 1$

So equation (6) has one solution.

Diophantine equation corresponding to (6):

$7y - 13k = 11 \quad \ldots(7)$

$13 = 7(1) + 6$

$7 = 6(1) + 1$

$6 = 1(6) + 0$

$1 = 7 - 6$

$= 7 - (13 - 7)$

$= 2\cdot 7 - 13$

Multiplying by 11:

$11 = 7(22) - 13(11) \quad \ldots(8)$

Comparing (7) and (8),

$y_0 = 22$, $k_0 = 11$

$y = y_0 + (m/d)t$, $t = 0$

$y = 22$

Thus $x = -6$ and $y = 22$ is a solution; equivalently, the unique solution modulo 13 is $x \equiv 7$ and $y \equiv 9 \pmod{13}$.

2) Solve

$7x + 3y \equiv 10 \pmod{16} \quad \ldots(1)$

$2x + 5y \equiv 7 \pmod{16} \quad \ldots(2)$

Comparing equation (1) with $ax + by \equiv r \pmod{m}$:

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

Comparing equation (2) with $cx + dy \equiv s \pmod{m}$:

$c = 2$, $d = 5$, $s = 7$, $m = 16$

$ad - bc = 7\times 5 - 3\times 2 = 35 - 6 = 29$

$$\gcd(ad - bc, m) = \gcd(29,16) = 1$$

Therefore the given system has a unique solution.

Eliminating y: multiply equation (1) by 5 and equation (2) by 3.

$$35x + 15y \equiv 50 \pmod{16}$$

$$6x + 15y \equiv 21 \pmod{16}$$

$$29x \equiv 29 \pmod{16}$$

This is a linear congruence in one variable.

$\gcd(29,16) = 1$

Hence it has one solution.

Diophantine equation:

$$29x - 16y = 29 \quad \ldots(4)$$

$29 = 16(1) + 13$

$16 = 13(1) + 3$

$13 = 3(4) + 1$

$3 = 1(3) + 0$

$$\begin{aligned} 1 &= 13 - 3\times4 \\ &= 13 - (16 - 13)\times4 \\ &= 13\times5 - 16\times4 \\ &= (29 - 16)\times5 - 16\times4 \\ &= 29\times5 - 16\times9 \end{aligned}$$

Multiplying by 29:

$$29\times5(29) - 16(29\times9) = 29$$

$$29(145) - 16(261) = 29$$

$x_0 = 145$, $y_0 = 261$

Let the initial value be $x_0 = 145$.

$x = x_0 + (m/d)t$, $t = 0$

$x = 145$

Eliminating x: multiply equation (1) by 2 and equation (2) by 7.

$$14x + 6y \equiv 20 \pmod{16}$$

$$14x + 35y \equiv 49 \pmod{16}$$

$$-29y \equiv -29 \pmod{16}$$

$$29y \equiv 29 \pmod{16} \quad \ldots(5)$$

$\gcd(29,16) = 1$

So equation (5) has one solution.

Diophantine equation:

$$29y - 16k = 29 \quad \ldots(6)$$

Multiplying the identity $29\times5 - 16\times9 = 1$ by 29:

$$29(145) - 16(261) = 29$$

$y_0 = 145$, $k_0 = 261$

$y = y_0 + (m/d)t$, where $t = 0$

$y = 145$

Therefore $x = 145$ and $y = 145$ is a solution; equivalently, the unique solution modulo 16 is $x \equiv 1$ and $y \equiv 1 \pmod{16}$.

**Simultaneous linear congruences**

A system of two or more linear congruences need not have a solution even though each individual congruence has a solution.

For example, there is no x which simultaneously satisfies $x \equiv 1 \pmod 2$ and $x \equiv 0 \pmod 4$, even though each congruence separately has a solution.

In the above example, moduli 2 and 4 are not relatively prime.

We shall prove that a system of two or more linear congruences can be solved simultaneously if the moduli are relatively prime in pairs.

**Chinese Remainder Theorem**

Assume $m_1, m_2, \ldots, m_r$ are positive integers and pairwise relatively prime; i.e.

$$\gcd(m_i, m_j) = 1,\ i \neq j$$

Let $a_1, a_2, \ldots, a_r$ be arbitrary integers. Then the system

$$x \equiv a_1 \pmod{m_1}$$
$$x \equiv a_2 \pmod{m_2}$$
$$\vdots$$
$$x \equiv a_r \pmod{m_r}$$

has a unique simultaneous solution modulo $m_1 m_2 \cdots m_r$.

**PROBLEMS**

---

**1) Using Chinese Remainder Theorem, solve the system of linear congruences**

$$x \equiv 2 \pmod 3 \quad \ldots(1)$$
$$x \equiv 3 \pmod 5 \quad \ldots(2)$$
$$x \equiv 2 \pmod 7 \quad \ldots(3)$$

Here $a_1 = 2$, $a_2 = 3$, $a_3 = 2$; $m_1 = 3$, $m_2 = 5$, $m_3 = 7$.

The moduli are pairwise relatively prime:

$$\gcd(3,5)=1,\ \gcd(3,7)=1,\ \gcd(5,7)=1$$
$$m = m_1 m_2 m_3 = 3\times5\times7 = 105$$

Let $x = a_1N_1x_1 + a_2N_2x_2 + a_3N_3x_3$, where $N_i = m/m_i$.

$$N_1 = 105/3 = 35$$
$$N_2 = 105/5 = 21$$
$$N_3 = 105/7 = 15$$

Now

$$\gcd(N_1,m_1) = \gcd(35,3)=1$$
$$35x_1 \equiv 1 \pmod 3$$

If $x_1 = 1$, then $35 \equiv 2 \pmod 3$, not 1.

If $x_1 = 2$, then $70 \equiv 1 \pmod 3$. Hence $x_1 = 2$.

$$\gcd(N_2,m_2) = \gcd(21,5)=1$$
$$21x_2 \equiv 1 \pmod 5$$

If $x_2 = 1$, $21 \equiv 1 \pmod 5$. Hence $x_2 = 1$.

$\gcd(N_3,m_3) = \gcd(15,7)=1$

$15x_3 \equiv 1 \pmod 7$

If $x_3 = 1$, $15 \equiv 1 \pmod 7$. Hence $x_3 = 1$.

$x = N_1a_1x_1 + N_2a_2x_2 + N_3a_3x_3$

$= (35)(2)(2) + (21)(3)(1) + (15)(2)(1)$

$= 140 + 63 + 30 = 233$

$x \equiv 233 \pmod{105}$

$\Rightarrow x \equiv 23 \pmod{105}$

2) Solve

$x \equiv 1 \pmod 4$, $x \equiv 0 \pmod 3$, $x \equiv 5 \pmod 7$

$m_1=4$, $m_2=3$, $m_3=7$; $a_1=1$, $a_2=0$, $a_3=5$

$m = 4\times3\times7 = 84$

$N_1 = 84/4 = 21$

$N_2 = 84/3 = 28$

$N_3 = 84/7 = 12$

$\gcd(N_1,m_1) = \gcd(21,4)=1$

$21x_1 \equiv 1 \pmod 4$

If $x_1 = 1$, $21 \equiv 1 \pmod 4$. Hence $x_1 = 1$.

$\gcd(N_2,m_2) = \gcd(28,3)=1$

$28x_2 \equiv 1 \pmod 3$

If $x_2 = 1$, $28 \equiv 1 \pmod 3$. Hence $x_2 = 1$.

$\gcd(N_3,m_3) = \gcd(12,7)=1$

$12x_3 \equiv 1 \pmod 7$

Testing values: $x_3 = 1$ gives 5, $x_3 = 2$ gives 3, $x_3 = 3$ gives 1 (mod 7). Hence $x_3 = 3$.

$x = N_1a_1x_1 + N_2a_2x_2 + N_3a_3x_3$

$= (21)(1)(1) + (28)(0)(1) + (12)(5)(3)$

$= 21 + 180 = 201$

$x \equiv 201 \pmod{84}$

$\Rightarrow x \equiv 33 \pmod{84}$

The smallest positive solution is $x = 33$.

3) Solve the following system of simultaneous linear congruences:

$x \equiv 2 \pmod 7$, $x \equiv 7 \pmod{16}$

Here $a_1 = 2$, $a_2 = 7$; $m_1 = 7$, $m_2 = 16$, and $\gcd(7,16)=1$.

$m = m_1m_2 = 7\times16 = 112$

$N_1 = m/m_1 = 112/7 = 16$

$N_2 = m/m_2 = 112/16 = 7$

$\gcd(N_1,m_1) = \gcd(16,7)=1$

$16x_1 \equiv 1 \pmod 7$

Testing: $16 \equiv 2 \pmod 7$. Thus $2x_1 \equiv 1 \pmod 7$, giving $x_1 = 4$.

$\gcd(N_2,m_2) = \gcd(7,16)=1$

$7x_2 \equiv 1 \pmod{16}$

Testing $x_2 = 1,2,\ldots$ gives $x_2 = 7$, since $7\times7 = 49 \equiv 1 \pmod{16}$.

$x = a_1N_1x_1 + a_2N_2x_2$

$= 2\times16\times4 + 7\times7\times7$

$= 128 + 343 = 471$

**$x \equiv 471 \pmod{112}$**

$\Rightarrow x \equiv 23 \pmod{112}$

## RANDOM VARIABLES

A random variable X is a real-valued function of the elements of the sample space of a random experiment. In other words, a variable which takes the real values, depending on the outcome of a random experiment is called a **random variable**, e.g.,

(i) When a fair coin is tossed, S = {H, T}. If X is the random variable denoting the number of heads,

$$X(H) = 1 \text{ and } X(T) = 0.$$

Hence, the random variable X can take values 0 and 1.

(ii) When two fair coins are tossed, S = {HH, HT, TH, TT}. If X is the random variable denoting the number of heads,

$$X(HH) = 2, X(HT) = 1, X(TH) = 1, X(TT) = 0.$$

Hence, the random variable X can take values 0, 1 and 2.

(iii) When a fair die is tossed, S = {1, 2, 3, 4, 5, 6}. If X is the random variable denoting the square of the number obtained,

$$X(1) = 1, X(2) = 4, X(3) = 9, X(4) = 16, X(5) = 25, X(6) = 36.$$

Hence, the random variable X can take values 1, 4, 9, 16, 25 and 36.

---

### Types of Random Variables

There are two types of random variables:

(i) **Discrete random variables** (ii) **Continuous random variables**

### Discrete Random Variables

A random variable X is said to be discrete if it takes either finite or countably infinite values. Thus, a discrete random variable takes only isolated values, e.g.,

(i) Number of children in a family

(ii) Number of cars sold by different companies in a year

(iii) Number of days of rainfall in a city

(iv) Number of stars in the sky

(v) Number of hot days in a year

**Continuous Random Variables**

A random variable X is said to be continuous if it takes any values in a given interval. Thus, a continuous random variable takes uncountably infinite values, e.g.,

(i) Height of a person in cm

(ii) Weight of a bag in kg

(iii) Temperature of a city in degree Celsius

(iv) Life of an electric bulb in hours

(v) Volume of a gas in cc

---

**Example 1: Identify the random variables as either discrete or continuous in each of the following cases:**

(i) A page in a book can have at most 300 words

**X = Number of misprints on a page**

(ii) Number of students present in a class of 50 students

(iii) A player goes to the gymnasium regularly

**X = Reduction in his weight in a month**

(iv) Number of attempts required by a candidate to clear the IAS examination

(v) Height of a skyscraper

---

**Solution**

(i) X = Number of misprints on a page.

The page may have no misprint or 1 misprint or 2 misprints ... or 300 misprints. Thus, X takes values 0, 1, 2, ... 300. Hence, X is a discrete random variable.

(ii) Let X be the random variable denoting the number of students present in a class. X takes values 0, 1, 2, ... 50. Hence, X is a discrete random variable.

(iii) Reduction in weight cannot take isolated values 0, 1, 2, etc., but it takes any continuous values. Hence, X is a continuous random variable.

(iv) Let X be a random variable denoting the number of attempts required by a candidate. Thus, X takes values 1, 2, 3, .... Hence, X is a discrete random variable.

(v) Since height can have any fractional value, it is a continuous random variable.

## PROBABILITY MASS FUNCTION

The probability distribution of a random variable is the set of its possible values together with their respective probabilities. Let X be a discrete random variable which takes the values $x_1$, $x_2$, ... $x_n$. The probability of each possible outcome $x_i$ is $p_i = p(x_i) = P(X = x_i)$ for i = 1, 2, ..., n. The number $p(x_i)$, i = 1, 2, ..., n must satisfy the following conditions:

(i) $p(x_i) \geq 0$ for all values of $x_i$

(ii) $\sum p(x_i) = 1$

The function $p(x_i)$ is called the probability function or probability mass function of the random variable X. The set of pairs $\{x_i, p(x_i)\}$, i = 1, 2, ..., n is called the probability distribution of X.

The **distribution of the random variable** which can be displayed in the form of a table is shown below:

| $X = x_i$ | $x_1$ | $x_2$ | $x_3$ | $\cdots$ | $\cdots$ | $x_n$ |
|---|---|---|---|---|---|---|
| $p(x_i) = P(X = x_i)$ | $p(x_1)$ | $p(x_2)$ | $p(x_3)$ | $\cdots$ | $\cdots$ | $p(x_n)$ |

## DISCRETE DISTRIBUTION FUNCTION

Let $X$ be a discrete random variable which takes the values $x_1, x_2, \ldots$ such that $x_1 \leq x_2 \leq \cdots$ with probabilities $p(x_1), p(x_2), \ldots$ such that $p(x_i) \geq 0$ for all values of $i$ and

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

The distribution function $F(x)$ of the discrete random variable $X$ is defined by

$$F(x) = P(X \leq x) = \sum_{x_i \leq x} p(x_i)$$

where $x$ is any integer. The function $F(x)$ is also called the cumulative distribution function. The set of pairs $[x_i, F(x_i)], i = 1,2, \ldots$, is called the cumulative probability distribution.

| X | $x_1$ | $x_2$ | ... |
|---|---|---|---|
| F(x) | $p(x_1)$ | $p(x_1) + p(x_2)$ | ... |

**Example 1: A fair die is tossed once. If the random variable is getting an even number, find the probability distribution of X.**

**Solution** When a fair die is tossed, $S = \{1,2,3,4,5,6\}$. Let $X$ be the random variable of getting an even number. Hence, $X$ can take the values 0 and 1.

$$P(X = 0) = P(1,3,5) = \frac{3}{6} = \frac{1}{2}$$

$$P(X = 1) = P(2,4,6) = \frac{3}{6} = \frac{1}{2}$$

Hence, the probability distribution of $X$ is:

| $X = x$ | 0 | 1 |
|---|---|---|
| $P(X = x)$ | ½ | ½ |

Also,

$$\sum P(X = x) = \frac{1}{2} + \frac{1}{2} = 1$$

---

**Example 2: Find the probability distribution of the number of heads when three coins are tossed.**

**Solution** When three coins are tossed, $S = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$

Let $X$ be the random variable of getting heads in tossing of three coins. Hence $X$ can take the values 0, 1, 2, 3.

$$P(X = 0) = P(no\ head) = P(TTT) = \frac{1}{8}$$

$$P(X = 1) = P(one\ head) = P(HTT, THT, TTH) = \frac{3}{8}$$

$$P(X = 2) = P(two\ heads) = P(HHT, HTH, THH) = \frac{3}{8}$$

$$P(X = 3) = P(three\ heads) = P(HHH) = \frac{1}{8}$$

Hence, the probability distribution of $X$ is:

| $X = x$ | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| $P(X = x)$ | ⅛ | ⅜ | ⅜ | ⅛ |

Also,

$$\sum P(X = x) = \frac{1}{8} + \frac{3}{8} + \frac{3}{8} + \frac{1}{8} = 1$$

---

**Example 3: State with reasons whether the following represent the probability mass function of a random variable:**

(i)

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

(ii)

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

(iii)

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

---

**Solution**

(i) Here, $0 \leq P(X = x) \leq 1$ is satisfied for all values of $X$.

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

Since $\sum P(X = x) = 1$, it represents probability mass function.

(ii) Here, $0 \leq P(X = x) \leq 1$ is satisfied for all values of $X$.

P(X = x) = P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) = $\frac{1}{2} + \frac{1}{2} + \frac{1}{2} + \frac{1}{4} = \frac{5}{4} > 1$

Since $\sum P(X = x) > 1$, it does not represent a probability mass function.

(iii) Here, $0 \leq P(X = x) \leq 1$ is **not** satisfied for all the values of $X$, as

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

Hence, $P(X = x)$ does not represent a probability mass function.

**Example 5:** A random variable $X$ has the probability mass function given by:

| $X$ | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| $P(X = x)$ | 0.1 | 0.2 | 0.5 | 0.2 |

Find (i) $P(2 \leq X < 4)$, (ii) $P(X > 2)$, (iii) $P(X\ is\ odd)$, and (iv) $P(X\ is\ even)$

**Solution**

(i) $P(2 \leq X < 4)$ = P(X = 2) + P(X = 3) ⇒ 0.2 + 0.5 = 0.7

(ii) $P(X > 2)$ = P(X = 3) + P(X = 4) ⇒ 0.5 + 0.2 = 0.7

(iii) $P(X\ is\ odd)$ = P(X = 1) + P(X = 3) ⇒ 0.1 + 0.5 = 0.6

(iv) $P(X\ is\ even)$ = P(X = 2) + P(X = 4) ⇒ 0.2 + 0.2 = 0.4

**Example 6:** If the random variable $X$ takes the values 1, 2, 3, and 4 such that

$$2P(X = 1) = 3P(X = 2) = 5P(X = 3) = 5P(X = 4)$$

Find the probability distribution.

**Solution** Let

$$2P(X = 1) = 3P(X = 2) = P(X = 3) = 5P(X = 4) = k$$

Then:

$$P(X = 1) = \frac{k}{2}, \quad P(X = 2) = \frac{k}{3}, \quad P(X = 3) = k, \quad P(X = 4) = \frac{k}{5}$$

Since

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

$$\Rightarrow \frac{k}{2} + \frac{k}{3} + k + \frac{k}{5} = 1$$

$$\Rightarrow k\left(\frac{1}{2} + \frac{1}{3} + k + \frac{1}{5}\right) = 1$$

$$\Rightarrow k\left(\frac{61}{30}\right) = 1 \Rightarrow k = \frac{30}{61}$$

Hence, the **probability distribution** is:

| $X = x$ | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| $P(X = x)$ | $\frac{15}{61}$ | $\frac{10}{61}$ | $\frac{30}{61}$ | $\frac{6}{61}$ |

---

**Example 10: Construct the distribution function of the discrete random variable X whose probability distribution is as given below:**

| $X$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| $P(X = x)$ | 0.1 | 0.15 | 0.25 | 0.2 | 0.15 | 0.1 | 0.05 |

---

**Solution: Distribution function of X:**

| $X$ | $P(X = x)$ | $F(x)$ |
|---|---|---|
| 1 | 0.1 | 0.1 |
| 2 | 0.15 | 0.25 |
| 3 | 0.25 | 0.5 |
| 4 | 0.2 | 0.7 |
| 5 | 0.15 | 0.85 |
| 6 | 0.1 | 0.95 |
| 7 | 0.05 | 1 |

---

**Example 11: A random variable X has the probability function given below:**

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

Find (i) $k$, (ii) $P(X < 2)$, (iii) $P(X \leq 2)$, (iv) $P(0 < X < 2)$, and (v) the distribution function.

---

**Solution**

(i) Since $P(X = x)$ is a probability mass function,

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

Hence, the probability distribution is:

| $X$ | 0 | 1 | 2 |
|---|---|---|---|
| $P(X = x)$ | $\frac{1}{6}$ | $\frac{2}{6}$ | $\frac{3}{6}$ |

---

**Example 13:** A discrete random variable $X$ has the following distribution function:

$$F(x) = \{0\ for\ x < 1\ \frac{1}{3}\ for\ 1 \leq x < 4\ \frac{1}{2}\ for\ 4 \leq x < 6\ \frac{5}{6}\ for\ 6 \leq x < 10\ 1\ for\ x \geq 10$$

Find: (i) $P(2 < X \leq 6)$, (ii) $P(X = 5)$, (iii) $P(X = 4)$, (iv) $P(X \leq 6)$, and (v) $P(X = 6)$

---

**Solution**

(i) $P(2 < X \leq 6) = F(6) - F(2) = \frac{5}{6} - \frac{1}{3} = \frac{3}{6} = \frac{1}{2}$

(ii) $P(X = 5) = P(X \leq 5) - P(X < 5) = F(5) - P(X < 5) = \frac{1}{2} - \frac{1}{2} = 0$

(iii) $P(X = 4) = P(X \leq 4) - P(X < 4) = F(4) - P(X < 4) = \frac{1}{2} - \frac{1}{3} = \frac{1}{6}$

(iv) $P(X \leq 6) = F(6) = \frac{5}{6}$

(v) $P(X = 6) = P(X \leq 6) - P(X < 6) = F(6) - P(X < 6) = \frac{5}{6} - \frac{1}{2} = \frac{1}{3}$

---

**PROBABILITY DENSITY FUNCTION**

Let $X$ be a continuous random variable such that the probability of the variable $X$ falling in the small interval

$$x - \frac{1}{2}dx \ to \ x + \frac{1}{2}dx$$

is $f(x)dx$, i.e.,

$$P\left(x - \frac{1}{2}dx \leq X \leq x + \frac{1}{2}dx\right) = f(x)dx$$

The function $f(x)$ is called the probability density function of the random variable $X$ and the continuous curve $y = f(x)$ is called the probability curve.

**Properties of Probability Density Function**

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

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

(iii) $P(a < X < b) = \int_{a}^{b} \quad f(x)dx$

**Note:** $P(X \geq x) = 1 - P(X < x)$

---

**CONTINUOUS DISTRIBUTION FUNCTION**

If $X$ is a continuous random variable having the probability density function $f(x)$, then the function

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} \quad f(t)dt, \quad -\infty < x < \infty$$

is called the distribution function or cumulative distribution function of the random variable $X$.

**Properties of Cumulative Distribution Function**

(i) $F(-\infty) = 0$

(ii) $F(\infty) = 1$

(iii) $0 \leq F(x) \leq 1, \quad -\infty < x < \infty$

(iv) $P(a < X < b) = F(b) - F(a)$

(v) $F'(x) = \frac{d}{dx}F(x) = f(x), \quad f(x) \geq 0$

---

**Problem1:** Show that the function $f(x)$ defined by

$$f(x) = \{\frac{1}{7}\ 1 < x < 8\ 0\ otherwise$$

is a probability density function for a random variable. Hence, find $P(3 < X < 10)$**.**

**Solution**

$$f(x) \geq 0 \quad in \quad 1 < x < 8$$

$$\int_{-\infty}^{\infty} f(x)dx = \int_{-\infty}^{1} f(x)dx + \int_{1}^{8} f(x)dx + \int_{8}^{\infty} f(x)dx$$

$$= 0 + \int_{1}^{8} \frac{1}{7}dx + 0$$

$$= \frac{1}{7}[x]_1^8$$

$$= \frac{1}{7}(8 - 1)$$

$$= \frac{7}{7} = 1$$

Hence, $f(x)$ is a probability density function.

$$P(3 < X < 10) = \int_{3}^{10} f(x)dx = \int_{3}^{8} f(x)dx + \int_{8}^{10} f(x)dx$$

$$= \int_{3}^{8} \frac{1}{7}dx + 0$$

$$= \frac{1}{7}[x]_3^8$$

$$= \frac{1}{7}(8 - 3)$$

$$= \frac{5}{7}$$

---

**Problem2:** Find the constant k such that the function

$$f(x) = \{kx^2\ 0 < x < 3\ 0\ otherwise$$

**is a probability density function and compute** (i) $P(1 < X < 2)$, (ii) $P(X < 2)$, and (iii) $P(X \geq 2)$

**Solution**

Since $f(x)$ is a probability density function,

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

$$\int_{-\infty}^{0} f(x)dx + \int_{0}^{3} f(x)dx + \int_{3}^{\infty} f(x)dx = 1$$

$$0 + \int_{0}^{3} kx^2dx + 0 = 1$$

$$k\left[\frac{x^3}{3}\right]_0^3 = 1$$

$$\frac{k}{3}(27 - 0) = 1$$

$$9k = 1 \Rightarrow k = \frac{1}{9}$$

Hence,

$$f(x) = \{\frac{1}{9}x^2\ 0 < x < 3\ 0\ otherwise$$

---

(i) $P(1 < X < 2) = \int_1^2 f(x)dx$

$$= \int_{1}^{2} \frac{1}{9}x^2dx$$

$$= \frac{1}{9}\left[\frac{x^3}{3}\right]_1^2$$

$$= \frac{1}{27}(8 - 1) = \frac{7}{27}$$

(ii) $P(X < 2) = \int_0^2 f(x)dx = \int_0^2 \frac{1}{9}x^2dx$

$$= \frac{1}{9}\left[\frac{x^3}{3}\right]_0^2 = \frac{1}{27}(8 - 0) = \frac{8}{27}$$

(iii) $P(X \geq 2) = 1 - P(X < 2) = 1 - \frac{8}{27} = \frac{19}{27}$

**Problem3:** If the probability density function of a random variable is given by

$$f(x) = k(1 - x^2) \quad 0 < x < 1$$
$$= 0 \quad otherwise$$

Find the value of k and the probabilities that a random variable having this probability density will take on a value (i) between 0.1 and 0.2, and (ii) greater than 0.5.

**Solution**

Since $f(x)$ is a probability density function,

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

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

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

$$k\left[x - \frac{x^3}{3}\right]_0^1 = 1$$

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

$$k \cdot \frac{2}{3} = 1 \Rightarrow k = \frac{3}{2}$$

Hence,

$$f(x) = \frac{3}{2}(1 - x^2) \quad 0 < x < 1$$
$$= 0 \quad otherwise$$

---

(i) Probability that the variable will take on a value between 0.1 and 0.2

$$P(0.1 < X < 0.2) = \int_{0.1}^{0.2} f(x)\,dx$$

$$= \int_{0.1}^{0.2} \frac{3}{2}(1 - x^2)dx$$

$$= \frac{3}{2}\left[x - \frac{x^3}{3}\right]_{0.1}^{0.2}$$

$$= \frac{3}{2}\left(\left[0.2 - \frac{0.008}{3}\right] - \left[0.1 - \frac{0.001}{3}\right]\right)$$

$$= 0.1465$$

---

(ii) **Probability that the variable will take on a value greater than 0.5**

$$P(X > 0.5) = \int_{0.5}^{\infty} f(x)\,dx$$

$$= \int_{0.5}^{1} f(x)\,dx + \int_{1}^{\infty} f(x)\,dx$$

$$= \int_{0.5}^{1} \frac{3}{2}(1 - x^2)\,dx + 0$$

$$= \frac{3}{2}\int_{0.5}^{1} (1 - x^2)\,dx$$

$$= \frac{3}{2}\left[x - \frac{x^3}{3}\right]_{0.5}^{1}$$

$$= \frac{3}{2}\left[\left(1 - \frac{1}{3}\right) - \left(0.5 - \frac{0.125}{3}\right)\right]$$

$$= 0.3125$$

---

**Problem4:** The length of time (in minutes) that a certain lady speaks on the telephone is found to be a random phenomenon, with a probability function specified by the function

$$f(x) = Ae^{-\frac{x}{5}}, \quad x \geq 0$$

$$= 0 \quad otherwise$$

**(i)** Find the value of $A$ that makes $f(x)$ a probability density function. **(ii)** What is the probability that the number of minutes that she will take over the phone is more than 10 minutes?

---

**Solution**

**(i)** For $f(x)$ to be a probability density function,

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

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

$$0 + \int_{0}^{\infty} Ae^{-\frac{x}{5}}dx = 1$$

$$A\left[-5e^{-\frac{x}{5}}\right]_0^{\infty} = 1$$

$$-5A(e^{-\infty} - e^{0}) = 1$$

$$-5A(0 - 1) = 1$$

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

Hence,

$$f(x) = \frac{1}{5}e^{-\frac{x}{5}}, \quad x \geq 0$$

$$= 0 \quad otherwise$$

---

**(ii)** $P(X > 10) = \int_{10}^{\infty} f(x)\,dx$

$$= \int_{10}^{\infty} \frac{1}{5}e^{-\frac{x}{5}}dx$$

$$= \frac{1}{5}\left[-5e^{-\frac{x}{5}}\right]_{10}^{\infty}$$

$$= -(e^{-\infty} - e^{-2})$$

$$= -(0 - e^{-2}) = e^{-2}$$

---

**Problem5: Let the continuous random variable X have the probability density function**

$$f(x) = \{\frac{2}{x^3}\ 1 < x < \infty\ 0\ otherwise$$

**Find** $F(x)$.

**Solution:**

$$F(x) = \int_{-\infty}^{x} f(t)dt = \int_{1}^{x} \frac{2}{t^3}dt = 0 + \int_{1}^{x} 2t^{-3}dt = 0 + 2\left[\frac{t^{-2}}{-2}\right]_1^x = -t^{-2}|_1^x$$
$$= -\left(\frac{1}{x^2} - 1\right) = 1 - \frac{1}{x^2}$$

**Hence,**

$$F(x) = \{1 - \frac{1}{x^2}\ 1 < x < \infty\ 0\ otherwise$$

**Problem6: Verify that the function $F(x)$ is a distribution function.**

$$F(x) = \{0\ x < 0\ 1 - e^{-x/4}\ x \geq 0$$

**Also, find the probabilities $P(X \leq 4),\ P(X > 8),\ P(4 \leq X \leq 8)$**

**Solution:** For the function $F(x)$,

(i) $F(-\infty) = 0$

(ii) $F(\infty) = 1 - e^{-\infty} = 1 - 0 = 1$

(iii) $0 \leq F(x) \leq 1,\ for\ all\ x$

(iv) $F(x)$ is non-decreasing

If $f(x)$ is the corresponding probability density function,

$$f(x) = \frac{d}{dx}F(x) \Rightarrow f(x) = \{0\ x < 0\ \frac{1}{4}e^{-x/4}\ x \geq 0$$

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

$$= 0 + \int_{0}^{\infty} \frac{1}{4}e^{-x/4}dx$$

$$= -\frac{1}{4} \cdot 4e^{-x/4}|_0^{\infty}$$

$$= -[e^{-x/4}|_0^{\infty}$$

$$= -[0 - 1]$$

$$= 1$$

Now, compute the probabilities:

(i) $P(X \leq 4) = F(4)$

$$= 1 - e^{-1} = \frac{e-1}{e}$$

(ii) $P(X > 8) = 1 - P(X \leq 8) = 1 - F(8)$

$$= 1 - (1 - e^{-2}) = e^{-2} = \frac{1}{e^2}$$

(iii) $P(4 \leq X \leq 8) = F(8) - F(4)$

$$= (1 - e^{-2}) - (1 - e^{-1}) = e^{-1} - e^{-2} = \frac{1}{e} - \frac{1}{e^2} = \frac{e-1}{e^2}$$

---

**Problem7: The probability density function of a continuous random variable X is given by**

$$f(x) = \{ax\ 0 \leq x \leq 1\ a\ 1 \leq x \leq 2\ 3a - ax\ 2 \leq x \leq 3\ 0\ otherwise$$

**(i) Find the value of a, (ii) Find the cdf of X.**

---

**Solution:**

(i) Since $f(x)$ is a probability density function,

$$\int_{-\infty}^{\infty} f(x)dx = \int_{-\infty}^{0} f(x)dx + \int_{0}^{1} ax\,dx + \int_{1}^{2} a\,dx + \int_{2}^{3} (3a - ax)\,dx = 1$$

$$= 0 + \left[\frac{ax^2}{2}\right]_0^1 + [ax]_1^2 + \left[3ax - \frac{ax^2}{2}\right]_2^3$$

$$= a\left[\frac{1}{2} - 0\right] + a(2-1) + \left[\left(9a - \frac{9a}{2}\right) - (6a - 2a)\right]$$

$$= \frac{a}{2} + a + \left(\frac{9a}{2} - 4a\right) \Rightarrow \frac{a}{2} + a + \frac{a}{2} = 1 \Rightarrow 2a = 1 \Rightarrow a = \frac{1}{2}$$

---

(ii) $F(x) = \int_{-\infty}^{x} f(t)dt$

**For** $0 \leq x \leq 1$**:**

$$F(x) = \int_{-\infty}^{0} f(x)dx + \int_{0}^{x} f(x)dx = 0 + \int_{0}^{x} at\,dt = \left[\frac{at^2}{2}\right]_0^x = \frac{ax^2}{2}$$

---

**For** $1 \le x \le 2$**:**

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

$$F(x) = \int_{0}^{1} at\,dt + \int_{1}^{x} a\,dt = \left[\frac{at^2}{2}\right]_0^1 + [at]_1^x = \frac{a}{2} + a(x-1) = \frac{a}{2} + ax - a = ax - \frac{a}{2}$$

---

**For** $2 \le x \le 3$**:**

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

$$F(x) = 0 + \int_{0}^{1} at\,dt + \int_{1}^{2} a\,dt + \int_{2}^{x} (3a - at)\,dt$$

$$= \frac{a}{2} + a + \left[3at - \frac{at^2}{2}\right]_2^x = \frac{a}{2} + a + \left[3ax - \frac{ax^2}{2} - (6a - 2a)\right]$$

$$= \frac{a}{2} + a + 3ax - \frac{ax^2}{2} - 4a = 3ax - \frac{ax^2}{2} - \frac{5a}{2}$$

---

**Hence, the cumulative distribution function (cdf)** $F(x)$ is:

$$F(x) = \{\frac{ax^2}{2}\ 0 \le x \le 1\ ax - \frac{a}{2}\ 1 \le x \le 2\ 3ax - \frac{ax^2}{2} - \frac{5a}{2}\ 2 \le x \le 3$$

---

## MEASURES OF STATISTICS FOR DISCREATE RANDOM VARIABLES

### 1. Mean

The mean or average value (μ) of the probability distribution of a discrete random variable $X$ is called as **expectation** and is denoted by $E(X)$.

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

Where $p(x)$ is the probability mass function of the discrete random variable $X$. Expectation of any function $\phi(x)$ of a random variable $X$ is given by

$$E[\phi(x)] = \sum_{i=1}^{\infty} \phi(x_i) \cdot p(x_i) = \sum \phi(x) \cdot p(x)$$

**Some important results on expectation:**

(i) $E(k) = k$

(ii) $E(kX) = kE(X)$

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

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

(v) $E(X + Y) = E(X) + E(Y)$ Provided $E(X)$ and $E(Y)$ exist

(vi) $E(XY) = E(X) \cdot E(Y)$ if $X$ and $Y$ are two independent random variables.

**Note:** $E(X^2) = \sum x^2 . p(x)$

---

(i) $E(k) = k$

**Proof:**

$E(k) = \sum_x (k) \cdot p(x) = k \sum_x p(x) = k$ (since $\sum_x p(x) = 1$)

---

(ii) $E(kX) = kE(X)$

**Proof:**

$E(kX) = \sum_x (kx) \cdot p(x) = k \sum_x x . p(x) = k\ E(x)$ (since $E(x) = \sum_x x . p(x)$)

---

**(iii)** $E(X + k) = E(X) + k$

**Proof:**

Let $k$ be a constant. Then,

$$E(X+k) = \sum_x (x+k)\cdot p(x) = \sum_x x\cdot p(x) + \sum_x k\cdot p(x)$$

Now,

- $\sum_x x\cdot p(x) = E(X)$
- $\sum_x k\cdot p(x) = k\cdot \sum_x p(x) = k\cdot 1 = k$

So,

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

---

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

**Proof:**

Let $a$, $b$ be constants. Then,

$$E(aX+b) = \sum_x (ax+b)\cdot p(x) = \sum_x ax\cdot p(x) + \sum_x b\cdot p(x)$$

Now,

- $\sum_x ax\cdot p(x) = a\cdot \sum_x x\cdot p(x) = a\cdot E(X)$
- $\sum_x b\cdot p(x) = b\cdot \sum_x p(x) = b\cdot 1 = b$

So,

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

---

**(v)** $E(X+Y) = E(X) + E(Y), \textit{provided } E(X) \textit{ and } E(Y) \textit{ exist}$

**Proof (for jointly discrete random variables):**

$$E(X+Y) = \sum_x \sum_y (x+y)\cdot p(x,y) = \sum_x \sum_y x\cdot p(x,y) + \sum_x \sum_y y\cdot p(x,y)$$

Now,

- $\sum_x \sum_y x\cdot p(x,y) = \sum_x x\cdot \sum_y p(x,y) = \sum_x x\cdot p_X(x) = E(X)$
- $\sum_x \sum_y y\cdot p(x,y) = \sum_y y\cdot \sum_x p(x,y) = \sum_y y\cdot p_Y(y) = E(Y)$

So,

$$E(X+Y) = E(X) + E(Y)$$

---

**(vi)** $E(XY) = E(X) \cdot E(Y)$, *if X and Y are independent*

**Proof:**

By definition,

$$E(XY) = \sum_x \sum_y xy \cdot p(x,y)$$

If $X$ and $Y$ are independent, then:

$$p(x,y) = p_X(x) \cdot p_Y(y) \Rightarrow E(XY) = \sum_x \sum_y xy \cdot p_X(x) \cdot p_Y(y)$$

Now rearrange the double sum:

$$E(XY) = \left(\sum_x x \cdot p_X(x)\right) \cdot \left(\sum_y y \cdot p_Y(y)\right) = E(X) \cdot E(Y)$$

---

**2. Median**

The median is the point which divides the entire distribution into two equal parts. If $X$ is a random variable, the value of $X = x$ for which the cumulative distribution function $F(x) = \frac{1}{2}$ is called the **median of X**.

For a discrete random variable $X$, if there exists no $x$ such that $F(x) = \frac{1}{2}$, then the **median M** of the probability distribution is given by

$$M = \frac{x_k + x_{k+1}}{2}$$

Where $F(x_k) < \frac{1}{2}$ and $F(x_{k+1}) > \frac{1}{2}$, and $x_k$ and $x_{k+1}$ are two consecutive values of $X$.

---

**3. Mode**

The **mode** is the value of discrete random variable $X$ for which the probability is **maximum**.

---

### 4. Mean Deviation

**Mean deviation** of the probability distribution of a discrete random variable $X$ is given by

$$MD = E\{|X - \mu|\}$$

$$= \sum_{i=1}^{\infty} |x_i - \mu| \cdot p(x_i)$$

$$= \sum |x - \mu| \cdot p(x)$$

Where $p(x)$ is the probability mass function of the discrete random variable $X$.

---

### 5. Variance

**Variance** characterizes the **variability** in the distributions. Even two distributions with the same mean can still have different dispersion of data about their means.

Variance of the probability distribution of a discrete random variable $X$ is given by:

$$Var(X) = \sigma^2 = E(X - \mu)^2$$

$$= E(X^2 - 2XE(X) + \mu^2)$$

$$= E(X^2) - E(2X\mu) + E(\mu^2)$$

$$= E(X^2) - 2\mu E(X) + \mu^2 \quad [\because \mu \; is \; constant]$$

$$= E(X^2) - 2\mu\mu + \mu^2 \;\; [\because \; E(X) = \; E(X)$$

$$= E(X^2) - \mu^2$$

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

---

**Some important results on variance:**

(i) $Var(k) = 0$

(ii) $Var(kX) = k^2 \cdot Var(X)$

(iii) $Var(X + k) = Var(X)$

(iv) $Var(aX + b) = a^2 \cdot Var(X)$

---

**Definition of variance in terms of expectation:** $Var(X) = E(X^2) - [E(X)]^2$

**(i)** $Var(k) = 0$

Let $X = k$, a constant.

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

So:

$$Var(X) = E(X^2) - [E(X)]^2 = k^2 - k^2 = 0$$

---

**(ii)** $Var(kX) = k^2 \cdot Var(X)$

Let $Y = kX$

- $E(Y) = E(kX) = kE(X)$
- $E(Y^2) = E[(kX)^2] = E[k^2X^2] = k^2E(X^2)$

Then:

$$Var(kX) = E(Y^2) - [E(Y)]^2 = k^2E(X^2) - [kE(X)]^2 = k^2(E(X^2) - [E(X)]^2) = k^2 \cdot Var(X)$$

---

**(iii)** $Var(X + k) = Var(X)$

Let $Y = X + k$

- $E(Y) = E(X + k) = E(X) + k$
- $E(Y^2) = E[(X + k)^2] = E[X^2 + 2kX + k^2] = E(X^2) + 2kE(X) + k^2$

Now apply variance formula:

$$Var(X + k) = E(Y^2) - [E(Y)]^2$$

Substitute:

$$= (E(X^2) + 2kE(X) + k^2) - (E(X) + k)^2$$
$$= E(X^2) + 2kE(X) + k^2 - (E(X)^2 + 2kE(X) + k^2)$$

Cancel terms:

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

---

**(iv)** $Var(aX + b) = a^2 \cdot Var(X)$

Let $Y = aX + b$

- $E(Y) = aE(X) + b$

- $E(Y^2) = E[(aX + b)^2] = E[a^2X^2 + 2abX + b^2] = a^2E(X^2) + 2abE(X) + b^2$

So:

$$Var(aX + b) = E(Y^2) - [E(Y)]^2$$

Substitute:

$$= (a^2E(X^2) + 2abE(X) + b^2) - (aE(X) + b)^2$$
$$= a^2E(X^2) + 2abE(X) + b^2 - (a^2E(X)^2 + 2abE(X) + b^2)$$

Cancel:

$$= a^2(E(X^2) - E(X)^2) = a^2 \cdot Var(X)$$

---

**6. Standard Deviation**

**Standard deviation** is the positive square root of the arithmetic mean of the squares of the deviations of the given values from their arithmetic mean. It is denoted by the Greek letter **σ**.

$$SD = \sigma = \sqrt{\sum_{i=1}^{\infty} x_i^2 \cdot p(x_i) - \mu^2}$$

$$= \sqrt{E(X^2) - \mu^2}$$

$$= \sqrt{E(X^2) - [E(X)]^2}$$

---

**Problem1:** *A random variable X has the following distribution:*

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

**Find (i) mean, (ii) variance, and (iii)** $P(1 < X < 6)$**.**

---

**Solution**

**(i) Mean**

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

$$= \left(1.\frac{1}{36}\right) + \left(2.\frac{3}{36}\right) + \left(3.\frac{5}{36}\right) + \left(4.\frac{7}{36}\right) + \left(5.\frac{9}{36}\right) + \left(6.\frac{11}{36}\right)$$

$$= \frac{161}{36} = 4.47$$

**(ii) Variance**

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

$$= \left(1^2.\ \frac{1}{36}\right) + \left(2^2.\frac{3}{36}\right) + \left(3^2.\frac{5}{36}\right) + \left(4^2.\frac{7}{36}\right) + \left(5^2.\frac{9}{36}\right) + \left(6^2.\frac{11}{36}\right) - (4.47)^2$$

$$= \frac{791}{36} - 19.98 = 1.99$$

**(iii)** $P(1 < X < 6) = P(X = 2) + P(X = 3) + P(X = 4) + P(X = 5)$

$$= \frac{3}{36} + \frac{5}{36} + \frac{7}{36} + \frac{9}{36} = \frac{24}{36}$$

$$= 0.67$$

**Problem2:** *The probability distribution of a random variable $X$ is given below.*

*Find* (i) $E(X)$, (ii) $Var(X)$, (iii) $E(2X - 3)$, and (iv) $Var(2X - 3)$

| X | –2 | –1 | 0 | 1 | 2 |
|---|---|---|---|---|---|
| P(X = x) | 0.2 | 0.1 | 0.3 | 0.3 | 0.1 |

**Solution**

**(i)** $E(X) = \sum x \cdot p(x)$

$$= (-2)(0.2) + (-1)(0.1) + 0(0.3) + 1(0.3) + 2(0.1) = 0$$

**(ii)** $Var(X) = \sum x^2 p(x) - [E(X)]^2$

$$= 4(0.2) + 1(0.1) + 0 + 1(0.3) + 4(0.1) - 0 = 1.6$$

**(iii)** $E(2X - 3) = 2E(X) - 3 = 2(0) - 3 = -3$

**(iv)** $Var(2X - 3) = 2^2 \cdot Var(X) = 4 \cdot (1.6) = 6.4$

**Problem3:** The probability distribution of a random variable $X$ is given by

| $X$ | –3 | 6 | 9 |
|---|---|---|---|
| $f(x)$ | 1/6 | 1/2 | 1/3 |

Compute $E(X)$, $E(X)^2$ and hence find $E(3X+2)^2$

**Solution:** $E(X) = \sum_{\forall_i} \; x_i f(x_i) = (-3)\frac{1}{6} + 6(\frac{1}{2}) + 9(\frac{1}{3}) = \frac{11}{2}$

$$E(X^2) = \sum_{\forall_i} \; x_i^2 f(x_i) = (-3)^2 \cdot \frac{1}{6} + 6^2(\frac{1}{2}) + 9^2(\frac{1}{3}) = \frac{93}{2}$$

$$E(3X+2)^2 = E[9x^2 + 12x + 4] = 9E(x)^2 + 12E(x) + 4$$

$$= [9(\frac{93}{2}) + 12(\frac{11}{2}) + 4] = \frac{837}{2} + \frac{132}{2} + 4$$

$$= \frac{977}{2} = 488.50$$

---

**Problem4:** *A random variable X has the following probability function:*

| x | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|---|
| p(x) | 0 | k | 2k | 2k | 3k | $k^2$ | $2k^2$ | $7k^2 + k$ |

---

**(i)** Determine **k**.

**(ii)** Evaluate $P(X < 6), P(X \geq 6), P(0 < X < 5)$, and $P(0 \leq X \leq 4)$.

**(iii)** If $p(x \leq K) > 0.5$, find the minimum value of K

**(iv)** Determine the **distribution function** of $X$.

**(v)** Find the **mean**.

**(vi)** Find the **variance**.

---

**Solution**

**(i)** Since $p(x)$ is a probability mass function,

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

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

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

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

$$k=\frac{1}{10} \quad or \quad k=-1$$

$$k=\frac{1}{10} \quad [\,Since\;\; p(x)\geq 0, x\in Z^+\,]$$

---

**(ii)**

$$P(X<6)=P(X=0)+P(X=1)+P(X=2)+P(X=3)+P(X=4)+P(X=5)$$

$$=0+0.1+0.2+0.2+0.3+0.01=0.81$$

$$P(X\geq 6)=1-P(X<6)=1-0.81=0.19$$

$$P(0<X<5)=P(X=1)+P(X=2)+P(X=3)+P(X=4)$$

$$=0.1+0.2+0.2+0.3=0.8$$

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

$$=0+0.1+0.2+0.2+0.3=0.8$$

---

(iii) The required minimum value of K is obtained as below

p(x ≤ 1 ) = p( x = 0) + p( x = 1 )  = 0.1

p(x ≤ 2 ) = p( x = 0) + p( x = 1 ) + p( x = 2)  = 0.3

p(x ≤ 3 ) = p( x = 0) + p( x = 1 ) + p( x = 2) + p( x = 3 )  = 0.5

p(x ≤ 4) = p( x = 0) + p( x = 1 ) + p( x = 2) + + p( x = 3 ) + p( x = 4 ) = 0.8 >0.5

The minimum value of K for which p (x ≤ K) > 0.5 is K = 4

**(iv)** Distribution function of *X*:

| x | p(x) | F(x) |
|---|---|---|
| 0 | 0 | 0 |
| 1 | 0.1 | 0.1 |

| x | p(x) | F(x) |
|---|---|---|
| 2 | 0.2 | 0.3 |
| 3 | 0.2 | 0.5 |
| 4 | 0.3 | 0.8 |
| 5 | 0.01 | 0.81 |
| 6 | 0.02 | 0.83 |
| 7 | 0.17 | 1.00 |

---

**(v)**

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

$$= 0 \cdot 0 + 1(0.1) + 2(0.2) + 3(0.2) + 4(0.3) + 5(0.01) + 6(0.02) + 7(0.17)$$

$$= 3.66$$

---

**(vi)**

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

$$= 0 + 1(0.1) + 4(0.2) + 9(0.2) + 16(0.3) + 25(0.01) + 36(0.02) + 49(0.17) - (3.66)^2$$

$$= 3.4044$$

---

**Problem5:** *A fair dice is tossed. Let the random variable $X$ denote the twice the number appearing on the dice. Write the probability distribution of $X$. Calculate mean and variance.*

---

**Solution**

Let $x$ be the random variable which denotes twice the number appearing on the dice.

**(i)** Probability distribution of $X$:

| x | 2 | 4 | 6 | 8 | 10 | 12 |
|---|---|---|---|---|---|---|
| p(x) | 1/6 | 1/6 | 1/6 | 1/6 | 1/6 | 1/6 |

---

**(ii) Mean:**

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

$$= 2\left(\frac{1}{6}\right) + 4\left(\frac{1}{6}\right) + 6\left(\frac{1}{6}\right) + 8\left(\frac{1}{6}\right) + 10\left(\frac{1}{6}\right) + 12\left(\frac{1}{6}\right)$$

$$= 7$$

---

**(iii) Variance:**

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

$$= 4\left(\frac{1}{6}\right) + 16\left(\frac{1}{6}\right) + 36\left(\frac{1}{6}\right) + 64\left(\frac{1}{6}\right) + 100\left(\frac{1}{6}\right) + 144\left(\frac{1}{6}\right) - (7)^2$$

$$= 11.67$$

---

**Problem6:** *Two unbiased dice are thrown at random. Find the probability distribution of the sum of the numbers on them. Also, find mean and variance.*

---

**Solution**

If a pair of fair dice is rolled, then

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

$= \{(1,1)(1,2)(1,3)(1,4)(1,5)(1,6); (2,1)(2,2)(2,3)(2,4)(2,5)(2,6)\ (3,1)(3,2)(3,3)(3,4)(3,5)(3,6); (4,1)(4,2$

Let $X$ be the random variable defined as the sum of two fair dice, then

$P(X = 2) = P(1,1) = 1/36$

$P(X = 3) = P[(1,2), (2,1)] = 2/36$

$P(X = 4) = P[(1,3), (2,2)(3,1)] = 3/36$

$P(X = 5) = P[(1,4), (2,3), (3,2), (4,1)] = 4/36$

$P(X = 6) = P[(1,5), (2,4), (3,3), (4,2), (5,1)] = 5/36$

$P(X = 7) = P[(1,6), (2,5), (3,4)(4,3), (5,2), (6,1)] = 6/36$

$P(X = 8) = P[(2,6, (3,5), (4,4), (5,3), (6,2)] = 5/36$

$P(X = 9) = P[(3,6), (4,5), (5,4), (6,3)] = 4/36$

$P(X = 10) = P[(4,6), (5,5), (6,4)] = 3/36$

$P(X = 11) = P[(5,6), (6,5)] = 2/36$

$P(X=12)=P[(6,6)]=1/36$

Hence the probability distribution of $X$ is given as follows:

| X | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| P(X) | 1/36 | 2/36 | 3/36 | 4/36 | 5/36 | 6/36 | 5/36 | 4/36 | 3/36 | 2/36 | 1/36 |

---

**Mean:**

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

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

$$=7$$

**Variance:**

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

$$=\frac{1}{36}(2^2+2\cdot 3^2+3\cdot 4^2+4\cdot 5^2+5\cdot 6^2+6\cdot 7^2+5\cdot 8^2+4\cdot 9^2+3\cdot 10^2+2\cdot 11^2+1\cdot 12^2)-7^2$$

$$=5.83$$

---

**Problem7:** *A sample of 3 items is selected at random from a box containing 10 items of which 4 are defective. Find the expected number of defective items.*

---

**Solution**

Let $X$ be the random variable which denotes the defective items.

- Total number of items = 10
- Number of good items = 6
- Number of defective items = 4

$$P(X=0)=P(no\ defective\ item)=\frac{{}^{6}C_3}{{}^{10}C_3}=\frac{1}{6}$$

$$P(X=1)=P(one\ defective\ item)=\frac{{}^{4}C_1\cdot{}^{6}C_2}{{}^{10}C_3}=\frac{1}{2}$$

$$P(X=2)=P(two\ defective\ items)=\frac{{}^{4}C_2\cdot{}^{6}C_1}{{}^{10}C_3}=\frac{3}{10}$$

$$P(X=3) = P(three\ defective\ items) = \frac{^{4}C_3}{^{10}C_3} = \frac{1}{30}$$

---

**Probability distribution:**

| X | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| P(X = x) | 1/6 | 1/2 | 3/10 | 1/30 |

---

**Expected number of defective items:**

$$E(X) = \sum xp(x)$$

$$= 0 + \left(\frac{1}{2}\right) + 2 \cdot \frac{3}{10} + 3 \cdot \frac{1}{30}$$

$$= 1.2$$

---

**Problem8:** *A player tosses two fair coins. He wins ₹100 if a head appears and ₹200 if two heads appear. On the other hand, he loses ₹500 if no head appears. Determine the expected value of the game. Is the game favorable to the player?*

---

**Solution**

Let $X$ be the random variable which denotes the number of heads appearing in tosses of two fair coins.

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

$$p(x_1) = P(X = 0) = P(no\ heads) = \frac{1}{4}$$

$$p(x_2) = P(X = 1) = P(one\ head) = \frac{2}{4} = \frac{1}{2}$$

$$p(x_3) = P(X = 2) = P(two\ heads) = \frac{1}{4}$$

- **Amount to be lost** if no head appears: $x_1 = -₹500$
- **Amount to be won** if one head appears: $x_2 = ₹100$
- **Amount to be won** if two heads appear: $x_3 = ₹200$

---

**Expected value of the game:**

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

$$= x_1p(x_1) + x_2p(x_2) + x_3p(x_3)$$

$$= -500\left(\frac{1}{4}\right) + 100\left(\frac{1}{2}\right) + 200\left(\frac{1}{4}\right)$$

$$= -₹25$$

---

**MEASURES OF STATISTICS FOR CONTINUOUS RANDOM VARIABLES**

1. **Mean** The mean or average value $\mu$ of the probability distribution of a continuous random variable $X$ is called the expectation and is denoted by $E(X)$.

$$\mu = E(X) = \int_{-\infty}^{\infty} xf(x)dx$$

Expectation of any function $\phi(x)$ of a continuous random variable $X$ is given by

$$E[\phi(X)] = \int_{-\infty}^{\infty} \phi(x)f(x)dx$$

---

2. **Median** The median is the point which divides the entire distribution into two equal parts. In case of a continuous distribution, the median is the point which divides the total area into two equal parts. Thus, if a continuous random variable $X$ is defined from $a$ to $b$ and $M$ is the median,

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

By solving any one of this equation, the median is obtained.

---

3. **Mode** The mode is value of $x$ for which $f(x)$ is maximum. Mode is given by

$$f'(x) = 0 \quad and \quad f''(x) < 0 \quad for \quad a < x < b$$

---

4. **Mean Deviation** Mean deviation of the probability distribution of a continuous random variable $X$ is given by

$$MD = \int_{a}^{b} |x - \mu| f(x)dx$$

---

5. **Variance** The variance of the probability distribution of a continuous random variable $X$ is given by

$$Var(X) = \sigma^2 = \int_a^b (x-\mu)^2 f(x)dx$$

$$= \int_a^b x^2 f(x)dx - \mu^2$$

---

6. **Standard Deviation** The standard deviation of the probability distribution of a continuous random variable $X$ is given by

$$SD = \sqrt{Var(X)} = \sigma$$

---

**Problem1.** For the continuous random variable having pdf

$$f(x) = 4x^3 \quad 0 \le x \le 1$$

$$= 0 \quad otherwise$$

Find the mean and variance of $X$.

---

**Solution**

Mean $= \mu = \int_{-\infty}^{\infty} xf(x)dx$

$$= \int_{-\infty}^{0} xf(x)dx + \int_0^1 xf(x)dx + \int_1^{\infty} xf(x)dx$$

$$= 0 + \int_0^1 x(4x^3)dx + 0$$

$$= 4\int_0^1 x^4 dx$$

$$= 4\left[\frac{x^5}{5}\right]_0^1$$

$$= 4\left(\frac{1}{5} - 0\right)$$

$$= \frac{4}{5}$$

---

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

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

$$= 0 + \int_{0}^{1} x^2(4x^3)dx + 0 - \left(\frac{4}{5}\right)^2$$

$$= 4\left[\frac{x^6}{6}\right]_0^1 - \frac{16}{25}$$

$$= 4\left(\frac{1}{6} - 0\right) - \frac{16}{25}$$

$$= \frac{4}{6} - \frac{16}{25}$$

$$= \frac{2}{3} - \frac{16}{25}$$

$$= \frac{50 - 48}{75}$$

$$= \frac{2}{75}$$

---

**Problem2.** For the triangular distribution

$$f(x) = x, \quad 0 < x \leq 1$$

$$= 2 - x, \quad 1 \leq x \leq 2$$

$$= 0, \quad otherwise$$

Find the mean and variance.

---

**Solution**

$$\mu = \int_{-\infty}^{\infty} xf(x)dx$$

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

$$= 0 + \int_0^1 x^2 dx + \int_1^2 x(2-x)dx + 0$$

$$= \int_0^1 x^2 dx + \int_1^2 (2x - x^2)dx$$

$$= \left[\frac{x^3}{3}\right]_0^1 + \left[2.(\frac{x^2}{2}) - \frac{x^3}{3}\right]_1^2$$

$$= \left(\frac{1}{3} - 0\right) + \left[\left(4 - \frac{8}{3}\right) - \left(1 - \frac{1}{3}\right)\right]$$

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

$$= 1$$

---

**Problem3.** A continuous random variable has the probability density function

$$f(x) = kxe^{-\lambda x^2} \quad x \geq 0,\ \lambda > 0$$

$$= 0 \quad otherwise$$

Determine (i) $k$, (ii) mean, and (iii) variance.

---

**Solution**

Since $f(x)$ is a probability density function,

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

$$= \int_{-\infty}^{0} f(x)dx + \int_0^{\infty} f(x)dx = 1$$

$$= 0 + \int_0^{\infty} kxe^{-\lambda x^2} dx = 1$$

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

$$k\left[(0-0) - \left(0 - \frac{1}{\lambda^2}\right)\right] = 1$$

$$\frac{k}{\lambda^2} = 1$$

$$k = \lambda^2$$

Hence,

$$f(x) = \lambda^2 x e^{-\lambda x}, \quad x \geq 0,\ \lambda \neq 0$$
$$= 0 \quad otherwise$$

---

(ii) **Mean**

$$\mu = \int_{-\infty}^{\infty} \quad xf(x)dx$$

$$= \int_{-\infty}^{0} \quad xf(x)dx + \int_{0}^{\infty} \quad xf(x)dx$$

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

$$= \lambda^2 \int_{0}^{\infty} \quad x^2 e^{-\lambda x} dx$$

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

$$= \lambda^2 \left[(0 - 0 + 0) - \left(0 - 0 - \frac{2}{\lambda^3}\right)\right]$$

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

---

(iii) **Variance**

$$\sigma^2 = \int_{0}^{\infty} \quad x^2 f(x)dx - \mu^2$$

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

$$= \lambda^2 \left[(0 - 0 + 0 - 0) - \left(0 - 0 + 0 - \frac{6}{\lambda^4}\right)\right] - \frac{4}{\lambda^2}$$

$$= \frac{6}{\lambda^2} - \frac{4}{\lambda^2}$$

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

---

**Problem4.** A continuous random variable has probability density function

$$f(x) = 6(x - x^2), \quad 0 \le x \le 1$$

Find the (i) mean, (ii) variance, (iii) median, and (iv) mode.

---

**Solution**

(i)

$$\mu = \int_{-\infty}^{\infty} xf(x)dx$$

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

$$= 0 + \int_{0}^{1} x \cdot 6(x - x^2)dx + 0$$

$$= 6\int_{0}^{1} (x^2 - x^3)dx$$

$$= 6\left[\frac{x^3}{3} - \frac{x^4}{4}\right]_0^1$$

$$= 6\left(\frac{1}{3} - \frac{1}{4}\right)$$

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

---

(ii)

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

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

$$= 0 + \int_{0}^{1} x^2 \cdot 6(x - x^2)dx + 0 - \frac{1}{4}$$

$$= 6\int_{0}^{1} (x^3 - x^4)dx - \frac{1}{4}$$

$$= 6\left[\frac{x^4}{4} - \frac{x^5}{5}\right]_0^1 - \frac{1}{4}$$

$$= 6\left(\frac{1}{4} - \frac{1}{5}\right) - \frac{1}{4}$$

$$= 6\left(\frac{5-4}{20}\right) - \frac{1}{4}$$

$$= 6 \cdot (\frac{1}{20}) - \frac{1}{4}$$

$$= \frac{3}{10} - \frac{1}{4}$$

$$= \frac{12-10}{40} = \frac{2}{40} = \frac{1}{20}$$

---

(iii)

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

$$\int_{0}^{M} 6(x - x^2)dx = \frac{1}{2}$$

$$6\left[\frac{x^2}{2} - \frac{x^3}{3}\right]_0^M = \frac{1}{2}$$

$$6\left(\frac{M^2}{2} - \frac{M^3}{3}\right) = \frac{1}{2}$$

$$3M^2 - 2M^3 = \frac{1}{2}$$

$$4M^3 - 6M^2 + 1 = 0$$

$$(2M-1)(2M^2-2M-1)=0$$

$$M=\frac{1}{2} \quad or \quad M=\frac{1\pm\sqrt{3}}{2}$$

Since $M \in (0,1)$,

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

Hence, median $M=\frac{1}{2}$.

---

(iv) Mode is the value of $x$ for which $f(x)$ is maximum. For $f(x)$ to be maximum, $f'(x)=0$ and $f''(x)<0$.

$$f'(x)=0$$

$$6(1-2x)=0$$

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

$$f''(x)=-12$$

$$\text{At } x=\frac{1}{2}, f''(x)=-12<0$$

Hence, $f(x)$ is maximum at $x=\frac{1}{2}$

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

---

**Problem 5:** The daily consumption of electric power (in millions of $kW$-hours) is a R.V. having the P.D.F. $f(x)=\frac{1}{9}xe^{-x/3}, x>0$

$0, \quad x\le 0$

If the total production is 12 million $kW$-hours, determine the probability that there is power cut (shortage) on any given day.

**Solution:** Probability that the power consumed is between 0 to 12 is

$P(0\le x\le 12)=\int_0^{12} f(x)dx=\int_0^{12} \frac{1}{9}xe^{-x/3}dx=-\frac{x}{3}e^{-x/3}-e^{-x/3}|_0^{12}=1-5e^{-4}$

Power supply is inadequate if daily consumption exceeds 12 million $kW$, i.e.,

$P(x>12)=1-P(0\le x\le 12)=1-[1-5e^{-4}]=5e^{-4}=0.0\,915781$

**Problem 6:** The probability density function of a random variable $X$ is

$$f(x) = \frac{1}{2} sinx, \qquad 0 \leq x \leq \pi$$

$$= 0 \qquad otherwise$$

Find the mean, mode, and median of the distribution and also, find the probability between 0 and $\frac{\pi}{2}$.

**Solution**

(i)

$$\mu = \int_{-\infty}^{\infty} xf(x)dx$$

$$= \int_{-\infty}^{0} xf(x)dx + \int_{0}^{\pi} xf(x)dx + \int_{\pi}^{\infty} xf(x)dx$$

$$= 0 + \int_{0}^{\pi} x\left(\frac{1}{2} sinx\right) dx + 0$$

$$= \frac{1}{2}\int_{0}^{\pi} xsinx\, dx$$

$$= \frac{1}{2}[-xcosx + sinx]_0^{\pi}$$

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

(ii) Mode is the value of $x$ for which $f(x)$ is maximum. For $f(x)$ to be maximum, $f'(x) = 0$ and $f''(x) < 0$.

$$f'(x) = 0$$

$$cosx = 0$$

$$x = \frac{\pi}{2}$$

$$f''(x) = -\frac{1}{2} sinx$$

At $x = \frac{\pi}{2}$, $f''(x) = -\frac{1}{2} < 0$. Hence, mode $= \frac{\pi}{2}$.

(iii)

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

$$\int_{0}^{M} \frac{1}{2} sinx\, dx = \frac{1}{2}$$

$$\frac{1}{2}[-cosx]_{0}^{M} = \frac{1}{2}$$

$$-\frac{1}{2}(cosM - 1) = \frac{1}{2}$$

$$1 - cosM = 1$$

$$cosM = 0$$

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

Hence, median $M = \frac{\pi}{2}$.

(iv)

$$P\left(0 < x < \frac{\pi}{2}\right) = \int_{0}^{\pi/2} f(x)\, dx$$

$$= \int_{0}^{\pi/2} \frac{1}{2} sinx\, dx$$

$$= -\frac{1}{2}[cosx]_{0}^{\pi/2}$$

$$= -\frac{1}{2}(0 - 1) = \frac{1}{2}$$

---

**Problem 7:** The cumulative distribution function of a continuous random variable $X$ is

$$F(x) = \{1 - e^{-2x}, x \geq 0, 0, x < 0.$$

Find the (i) probability density function, (ii) mean, and (iii) variance.

---

**Solution**

(i) $f(x) = \frac{d}{dx}F(x)$

$$f(x) = \frac{1}{2}e^{-2x}, \quad x \geq 0$$

$$= 0, \quad x < 0$$

(ii) $\mu = \int_{-\infty}^{\infty} \quad xf(x)\,dx$

$$= \int_{0}^{\infty} \quad xf(x)\,dx$$

$$= \frac{1}{2}\int_{0}^{\infty} \quad xe^{-2x}\,dx$$

$$= \frac{1}{2}\left[x\left(\frac{e^{-2x}}{-2}\right) - 1\left(\frac{e^{-2x}}{4}\right)dx\right]_{0}^{\infty}$$

$$= \frac{1}{2}\left[(0-0) - \left(0 - \frac{1}{4}\right)\right]$$

$$= \frac{1}{8}$$

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

$$= \int_{0}^{\infty} \quad x^2 f(x)\,dx - \mu^2$$

$$= \frac{1}{2}\int_{0}^{\infty} \quad x^2 e^{-2x}\,dx - \frac{1}{64}$$

$$= \frac{1}{2}\left[x^2\left(\frac{e^{-2x}}{-2}\right) - 2x\left(\frac{e^{-2x}}{4}\right) + 2\left(\frac{e^{-2x}}{-8}\right)\right]_{0}^{\infty} - \frac{1}{64}$$

$$= \frac{1}{2}[(0-0-0)\left(0 - 0 - \frac{1}{4}\right) - \frac{1}{64}$$

$$= \frac{1}{8} - \frac{1}{64}$$

$$= \frac{7}{64}$$

## BINOMIAL DISTRIBUTION

Consider $n$ independent trials of a random experiment which results in either success or failure. Let $p$ be the probability of success (remaining constant every time) and $q = 1 - p$ be the probability of failure.

The probability of $x$ successes and $n - x$ failures is given by $p^x q^{n-x}$ (multiplication theorem of probability). But these $x$ successes and $n - x$ failures can occur in any one of the ${}^nC_x$ ways in each of which the probability is same. Hence, the probability of $x$ successes is

$$P(X = x) = {}^nC_x \cdot p^x \cdot q^{n-x}, \quad x = 0,1,2,\ldots,n, \quad where\, p + q = 1$$

A random variable $X$ is said to follow the binomial distribution if the probability of $x$ is given by

$$P(X = x) = p(x) = {}^nC_x \cdot p^x \cdot q^{n-x}, \quad x = 0,1,2,\ldots,n, \quad p + q = 1$$

The two constants $n$ and $p$ are called the parameters of the distribution.

---

### Examples of Binomial Distribution

(i) Number of defective bolts in a box containing $n$ bolts.

(ii) Number of post-graduates in a group of $n$ people.

(iii) Number of oil wells yielding natural gas in a group of $n$ wells tested.

(iv) Number of machines lying idle in a factory having $n$ machines.

---

### Conditions for Binomial Distribution

The binomial distribution holds under the following conditions:

(i) The number of trials $n$ is finite.

(ii) There are only two possible outcomes: success or failure.

(iii) The trials are independent of each other.

(iv) The probability of success $p$ is constant for each trial.

---

### Constants of the Binomial Distribution

---

### 1. Mean of the Binomial Distribution

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

$$= \sum_{x=0}^{n} x \cdot {}^{n}C_x \cdot p^x \cdot q^{n-x}$$

$$= 0 \cdot {}^{n}C_0 \cdot p^0 \cdot q^n + 1 \cdot {}^{n}C_1 \cdot p \cdot q^{n-1} + 2 \cdot {}^{n}C_2 \cdot p^2 \cdot q^{n-2} + \cdots + n \cdot {}^{n}C_n \cdot p^n$$

$$= nq^{n-1}p + 2.\frac{n(n.-1)}{2.1}q^{n-2}p^2 + 3.\frac{n(n-1)(n-2)}{3.2.1}q^{n-3}p^3 + \cdots + n.p^n$$

$$= nq^{n-1}p + n(n-1)q^{n-2}p^2 + \frac{n(n-1.)(n-2)}{2.1}q^{n-3}p^3 + \cdots + np^n$$

$$np[q^{n-1} + (n-1)q^{n-2}p^1 + \frac{(n-1.)(n-2)}{2.1}q^{n-3}p^2 + \ +np^{n-1}]$$

$$= np(q+p)^{n-1} = np \qquad (since\ p + q = 1)$$

$$\Rightarrow E(X) = np \quad [since\ p + q = 1]$$

---

**2. Variance of the Binomial Distribution**

$$Var(X) = E(X^2) - \mu^2$$

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

$$= \sum_{x=0}^{n} x^2 \cdot {}^{n}C_x \cdot p^x \cdot q^{n-x} - \mu^2$$

$$= \sum_{x=0}^{n} [x + x(x-1)] \cdot {}^{n}C_x \cdot p^x \cdot q^{n-x} - \mu^2$$

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

$$= np + [2.1.nC_2q^{n-2}p^2 + 3.2.nC_3q^{n-3}p^s + \ +n(n-1)^nC_np^n] - \mu^2$$

$$= np + [2.1.\frac{n(n.-1)}{2.1}q^{n-2}p^2 + 3.2.\frac{n(n-.1)(n-2)}{32.1}q^{n-3}p^3 + \ +n(n-1)p^n] - \mu^2$$

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

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

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

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

$$= np + n(n-1)p^2 - n^2p^2$$

$$= np + n^2p^2 - np^2 - n^2p^2$$

$$= np - np^2$$

$$= np[1-p] = npq \; [\because \; 1-p=q]$$

$$= npq$$

---

### 3. Standard Deviation of the Binomial Distribution

$$SD = \sqrt{Variance} = \sqrt{npq}$$

---

### 4. Mode of the Binomial Distribution

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

- Mode = integral part of $(n+1)p$, if $(n+1)p$ is not an integer
- Mode = $(n+1)p$ and $(n+1)p-1$, if $(n+1)p$ is an integer

---

### Recurrence Relation for the Binomial Distribution

For the binomial distribution,

$$P(x) = {}^nC_x \, p^x \, q^{n-x} = \frac{n!}{(n-x)!\,x!} \, p^x \, q^{n-x}$$

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

$$\frac{P(x+1)}{P(x)} = \frac{(n-x)!}{(n-x-1)!} \times \frac{x!}{(x+1)!} \times \frac{p}{q}$$

$$= \frac{(n-x) \times (n-x-1)!}{(n-x-1)!} \times \frac{x!}{(x+1) \times x!} \times \frac{p}{q} \times = (\frac{n-x}{x+1}) \cdot \frac{p}{q}$$

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

---

### Binomial Frequency Distribution

If $n$ independent trials constitute one experiment and this experiment is repeated $N$ times, the frequency of $x$ successes is $N \cdot P(X = x)$, i.e.,

$$N \cdot {}^{n}C_{x} \cdot p^{x} \cdot q^{n-x}$$

This is called **expected or theoretical frequency** $f(x)$ of a success.

$$\sum_{x=0}^{n} f(x) = N \sum_{x=0}^{n} P(X = x) = N \quad \left[ \because \sum_{x=0}^{n} P(X = x) = 1 \right]$$

The expected or theoretical frequencies $f(0), f(1), f(2), \dots, f(n)$ of $0,1,2,\dots,n$ successes are respectively the first, second, third, ..., $(n+1)^{th}$ term in the expansion of $N(p+q)^{n}$. The possible number of successes and their frequencies is called a binomial frequency distribution. In practice, the expected frequencies differ from observed frequencies due to chance factor.

---

**Problem1:**

**The mean and standard deviation of a binomial distribution are 5 and 2. Determine the distribution.**

---

**Solution**

$$\mu = np = 5$$

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

$$npq = 4$$

$$\Rightarrow \frac{npq}{np} = \frac{4}{5} \Rightarrow q = \frac{4}{5} \Rightarrow p = 1 - q = 1 - \frac{4}{5} = \frac{1}{5}$$

$$np = 5, \quad \left(\frac{1}{5}\right) n = 5 \Rightarrow n = 25$$

Hence, the binomial distribution is

$$P(X = x) = {}^{25}C_{x} \left(\frac{1}{5}\right)^{x} \left(\frac{4}{5}\right)^{25-x}, \quad x = 0,1,2,\dots,25$$

---

**Problem2:**

**The mean and variance of a binomial variate are 8 and 6. Find $P(X \geq 2)$.**

---

**Solution**

$$\mu = np = 8, \quad \sigma^2 = npq = 6$$

$$\frac{npq}{np} = \frac{6}{8} = \frac{3}{4} \Rightarrow q = \frac{3}{4} \Rightarrow p = 1 - q = 1 - \frac{3}{4} = \frac{1}{4}$$

$$np = 8, \quad n\left(\frac{1}{4}\right) = 8 \Rightarrow n = 32$$

$$P(X = x) = {}^{32}C_x \left(\frac{1}{4}\right)^x \left(\frac{3}{4}\right)^{32-x}, \quad x = 0,1,2,\ldots,32$$

$$P(X \geq 2) = 1 - P(X < 2)$$

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

$$= 1 - \sum_{x=0}^{1} P(X = x)$$

$$= 1 - \sum_{x=0}^{1} {}^{32}C_x \left(\frac{1}{4}\right)^x \left(\frac{3}{4}\right)^{32-x}$$

$$= 0.9988$$

---

**Problem3:**

**The mean and variance of a binomial distribution are 4 and $\frac{4}{3}$ respectively. Find$P(X \geq 1)$.**

---

**Solution**

$$\mu = np = 4$$

$$\sigma^2 = npq = \frac{4}{3}$$

$$\frac{npq}{np} = \frac{4/3}{4} = \frac{1}{3} \Rightarrow q = \frac{1}{3} \Rightarrow p = 1 - q = 1 - \frac{1}{3} = \frac{2}{3}$$

$$np = 4 \Rightarrow n\left(\frac{2}{3}\right) = 4 \Rightarrow n = 6$$

$$P(X = x) = {}^{6}C_x \left(\frac{2}{3}\right)^x \left(\frac{1}{3}\right)^{6-x}, \quad x = 0,1,2,\ldots,6$$

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

$$= 1 - {}^6C_0 \left(\frac{2}{3}\right)^0 \left(\frac{1}{3}\right)^6 = 1 - \left(\frac{1}{3}\right)^6 = 0.9986$$

---

**Problem4:**

**With the usual notation, find $p$ for a binomial distribution if $n = 6$ and $9P(X = 4) = P(X = 2)$.**

---

**Solution**

For the binomial distribution,

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

Given:

$$n = 6$$

$$9P(X = 4) = P(X = 2)$$

$$9 \cdot {}^6C_4 \cdot p^4 \cdot q^2 = {}^6C_2 \cdot p^2 \cdot q^4$$

$$9 \cdot {}^6C_4 \cdot p^4 \cdot q^2 = {}^6C_2 \cdot p^2 \cdot q^4 \Rightarrow 9p^2 = q^2 = (1 - p)^2$$

$$9p^2 = 1 - 2p + p^2 \Rightarrow 8p^2 + 2p - 1 = 0$$

Now solving the quadratic:

$$p = \frac{-2 \pm \sqrt{4 + 32}}{2 \cdot 8} = \frac{-2 \pm \sqrt{36}}{16} = \frac{-2 \pm 6}{16}$$

$$= \frac{1}{2}, \quad \frac{1}{4}$$

Since probability cannot be negative,

$$p = \frac{1}{4}$$

---

**Problem5:**

**If the probability of a defective bolt is $\frac{1}{8}$, find the (i) mean, and (ii) variance for the distribution of 640 defective bolts.**

## Solution

$$p = \frac{1}{8}, \quad n = 640$$

**(i)** Mean:

$$\mu = np = \frac{640}{8} = 80$$

**(ii)** Variance:

$$q = 1 - p = 1 - \frac{1}{8} = \frac{7}{8}$$

$$Variance = npq = 640 \cdot \left(\frac{1}{8}\right) \cdot \left(\frac{7}{8}\right) = 70$$

---

**Problem6:**

**4 coins are tossed simultaneously. What is the probability of getting (i) 2 heads? (ii) at least 2 heads? (iii) at most 2 heads?**

---

## Solution

Let $p$ be the probability of getting a head in the toss of a coin.

$$p = \frac{1}{2}, \quad q = 1 - p = \frac{1}{2}, \quad n = 4$$

The probability of getting $x$ heads when 4 coins are tossed is:

$$P(X = x) = {}^4C_x \cdot p^x \cdot q^{4-x} = {}^4C_x \left(\frac{1}{2}\right)^x \left(\frac{1}{2}\right)^{4-x}, \quad x = 0,1,2,3,4$$

---

**(i)** Probability of getting 2 heads:

$$P(X = 2) = {}^4C_2 \cdot \left(\frac{1}{2}\right)^2 \cdot \left(\frac{1}{2}\right)^2 = \frac{6}{16} = \frac{3}{8}$$

**(ii) Probability of getting at least two heads when 4 coins are tossed**

$$P(X \geq 2) = P(X = 2) + P(X = 3) + P(X = 4)$$

$$= \sum_{x=2}^{4} P(X = x) = \sum_{x=2}^{4} {}^{4}C_x \left(\frac{1}{2}\right)^x \left(\frac{1}{2}\right)^{4-x}$$

$$= \frac{11}{16}$$

---

**(iii) Probability of getting at most 2 heads when 4 coins are tossed**

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

$$= \sum_{x=0}^{2} P(X = x) = \sum_{x=0}^{2} {}^{4}C_x \left(\frac{1}{2}\right)^x \left(\frac{1}{2}\right)^{4-x} = \frac{11}{16}$$

---

### Problem7:

**Two dice are thrown five times. Find the probability of getting the sum as 7 (i) at least once, (ii) two times, and (iii)** $P(1 < X < 5)$

---

### Solution

In a single throw of two dice, a sum of 7 can occur in 6 ways out of $6 \times 6 = 36$ ways: (1,6), (6,1), (2,5), (5,2), (3,4), (4,3)

Let $p$ be the probability of getting the sum as 7 in a single throw:

$$p = \frac{6}{36} = \frac{1}{6}, \quad q = 1 - p = \frac{5}{6}, \quad n = 5$$

The binomial distribution:

$$P(X = x) = {}^{5}C_x \cdot p^x \cdot q^{5-x} = {}^{5}C_x \left(\frac{1}{6}\right)^x \left(\frac{5}{6}\right)^{5-x}, \quad x = 0,1,2,\ldots,5$$

**(i) Probability of getting the sum 7 at least once in 5 throws of two dice**

$$P(X \geq 1) = 1 - P(X = 0) = 1 - {}^{5}C_0 \left(\frac{1}{6}\right)^0 \left(\frac{5}{6}\right)^5 = 1 - \left(\frac{3125}{7776}\right) = \frac{4651}{7776}$$

---

**(ii) Probability of getting the sum 7 exactly two times in 5 throws of two dice**

$$P(X = 2) = {}^{5}C_2 \left(\frac{1}{6}\right)^2 \left(\frac{5}{6}\right)^3 = \frac{625}{3888}$$

---

**(iii) Probability of getting the sum 7 for $1 < X < 5$**

$$P(1 < X < 5) = P(X = 2) + P(X = 3) + P(X = 4)$$

$$= \sum_{x=2}^{4} {}^5C_x \left(\frac{1}{6}\right)^x \left(\frac{5}{6}\right)^{5-x} = \frac{1525}{7776}$$

---

**Problem8:**

**If 10% of the screws produced by a machine are defective, find the probability that out of 5 screws chosen at random, (i) none is defective, (ii) one is defective, and (iii) at most two are defective.**

---

**Solution**

Let $p = 0.1$, $q = 1 - p = 0.9$, $n = 5$

$$P(X = x) = {}^5C_x \cdot (0.1)^x \cdot (0.9)^{5-x}, \quad x = 0,1,\ldots,5$$

---

**(i) Probability that none of the 5 screws is defective**

$$P(X = 0) = {}^5C_0 \cdot (0.1)^0 \cdot (0.9)^5 = 0.59049$$

---

**(ii) Probability that one screw out of 5 screws is defective**

$$P(X = 1) = {}^5C_1 \cdot (0.1)^1 \cdot (0.9)^4 = 0.3281$$

---

**(iii) Probability that at most 2 screws are defective**

$$P(X \leq 2) = P(X = 0) + P(X = 1) + P(X = 2) = \sum_{x=0}^{2} P(X = x)$$

$$= \sum_{x=0}^{2} {}^5C_x \cdot (0.1)^x \cdot (0.9)^{5-x} = 0.99144$$

---

**Problem9:**

**The probability of a man hitting a target is $\frac{1}{3}$. (i) If he fires 5 times, what is the probability of his hitting the target at least twice? (ii) How many times must he fire so that the probability of his hitting the target at least once is more than 90%?**

---

**Solution**

Let $p$ be the probability of hitting the target:

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

Probability of hitting the target $x$ times out of 5 times:

$$P(X = x) = {}^5C_x \cdot p^x \cdot q^{5-x} = {}^5C_x \left(\frac{1}{3}\right)^x \left(\frac{2}{3}\right)^{5-x}, \quad x = 0,1,2,\ldots,5$$

---

**(i) Probability of hitting the target at least twice out of 5 times:**

$$P(X \geq 2) = P(X = 2) + P(X = 3) + P(X = 4) + P(X = 5)$$

$$= \sum_{x=2}^{5} P(X = x) = \sum_{x=2}^{5} {}^5C_x \left(\frac{1}{3}\right)^x \left(\frac{2}{3}\right)^{5-x}$$

$$= \frac{131}{243} \approx 0.5391$$

---

**(ii) Probability of hitting the target at least once out of 5 times:**

We want:

$$P(X \geq 1) > 0.9 \Rightarrow 1 - P(X = 0) > 0.9 \Rightarrow P(X = 0) < 0.1$$

$$P(X = 0) = {}^nC_0 \left(\frac{1}{3}\right)^0 \left(\frac{2}{3}\right)^n = \left(\frac{2}{3}\right)^n$$

We want:

$$1 - \left(\frac{2}{3}\right)^n > 0.9 \Rightarrow \left(\frac{2}{3}\right)^n < 0.1$$

Try $n = 6$:

$$\left(\frac{2}{3}\right)^6 = 0.088 \Rightarrow 1 - \left(\frac{2}{3}\right)^6 = 0.9122 > 0.9$$

Hence, the man must fire **at least 6 times** so that the probability of hitting the target at least once is more than 90%.

---

**Problem10:**

**Out of 800 families with 5 children each, how many would you expect to have (i) 3 boys? (ii) 5 girls? (iii) Either 2 or 3 boys? (iv) at least one boy? Assume equal probabilities for boys and girls.**

---

**Solution**

Let $p$ be the probability of having a boy in each family:

$$p = \frac{1}{2}, \quad q = 1 - p = \frac{1}{2}, \quad n = 5, \quad N = 800$$

The probability of having $x$ boys out of 5 children in each family:

$$P(X = x) = {}^5C_x \cdot p^x \cdot q^{5-x} = {}^5C_x \left(\frac{1}{2}\right)^x \left(\frac{1}{2}\right)^{5-x}, \quad x = 0,1,2,\ldots,5$$

---

**(i) Probability of having 3 boys out of 5 children in each family**

$$P(X = 3) = {}^5C_3 \left(\frac{1}{2}\right)^3 \left(\frac{1}{2}\right)^2 = \frac{5}{16}$$

$$\textit{Expected number of families having 3 boys out of 5 children} = N \cdot P(X = 3)$$
$$= 800 \cdot \frac{5}{16} = 250$$

---

**(ii) Probability of having 5 girls, i.e., no boys out of 5 children in each family**

$$P(X = 0) = {}^5C_0 \left(\frac{1}{2}\right)^0 \left(\frac{1}{2}\right)^5 = \frac{1}{32}$$

$$\textit{Expected number of families with 5 girls out of 5 children} = N \cdot P(X = 0) = 800 \cdot \frac{1}{32}$$
$$= 25$$

---

**(iii) Probability of having either 2 or 3 boys out of 5 children in each family**

$$P(X=2)+P(X=3)=\sum_{x=2}^{3} P(X=x)=\sum_{x=2}^{3} {}^{5}C_x\left(\frac{1}{2}\right)^x\left(\frac{1}{2}\right)^{5-x}=\frac{5}{8}$$

$$\textit{Expected number of families having either 2 or 3 boys out of 5 children} = N\cdot P(X=2\ or\ 3)=800\cdot\frac{5}{8}=500$$

---

**(iv) Probability of having at least one boy out of 5 children in each family**

$$P(X\geq 1)=P(X=1)+P(X=2)+P(X=3)+P(X=4)+P(X=5)=\sum_{x=1}^{5} P(X=x)$$
$$=\sum_{x=1}^{5} {}^{5}C_x\left(\frac{1}{2}\right)^x\left(\frac{1}{2}\right)^{5-x}=\frac{31}{32}$$

$$\textit{Expected number of families having at least one boy out of 5 children} = N\cdot P(X\geq 1)=800\cdot\frac{31}{32}=775$$

---

**Problem11:**

**Seven unbiased coins are tossed 128 times and the number of heads obtained is noted as given below:**

| No. of heads | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|---|
| Frequency | 7 | 6 | 19 | 35 | 30 | 23 | 7 | 1 |

**Fit a binomial distribution to the data.**

---

**Solution**

Since the coin is unbiased,

$$p=\frac{1}{2},\quad q=\frac{1}{2},\quad n=7,\quad N=128$$

For binomial distribution,

$$P(X=x)={}^{n}C_x\cdot p^x\cdot q^{n-x}={}^{7}C_x\left(\frac{1}{2}\right)^x\left(\frac{1}{2}\right)^{7-x},\quad x=0,1,2,\dots,7$$

Theoretical or expected frequency

$$f(x) = N \cdot P(X = x) = 128 \cdot {}^7C_x \left(\frac{1}{2}\right)^7$$

$$f(0) = 128 \cdot {}^7C_0 \left(\frac{1}{2}\right)^7 = 1$$

$$f(1) = 128 \cdot {}^7C_1 \left(\frac{1}{2}\right)^7 = 7$$

$$f(2) = 128 \cdot {}^7C_2 \left(\frac{1}{2}\right)^7 = 21$$

$$f(3) = 128 \cdot {}^7C_3 \left(\frac{1}{2}\right)^7 = 35$$

$$f(4) = 128 \cdot {}^7C_4 \left(\frac{1}{2}\right)^7 = 35$$

$$f(5) = 128 \cdot {}^7C_5 \left(\frac{1}{2}\right)^7 = 21$$

$$f(6) = 128 \cdot {}^7C_6 \left(\frac{1}{2}\right)^7 = 7$$

$$f(7) = 128 \cdot {}^7C_7 \left(\frac{1}{2}\right)^7 = 1$$

| Binomial distribution | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| No. of heads $x$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| Expected binomial frequency $f(x)$ | 1 | 7 | 21 | 35 | 35 | 21 | 7 | 1 |

---

**Problem12: Fit a binomial distribution to the following data:**

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

---

**Solution**

$$Mean = \frac{\sum fx}{\sum f} = \frac{2(0) + 14(1) + 20(2) + 34(3) + 22(4) + 8(5)}{2 + 14 + 20 + 34 + 22 + 8} = \frac{284}{100} = 2.84$$

For binomial distribution,

$$n = 5$$

**Given:**

$$\mu = np = 2.84, \quad n = 5$$

$$\Rightarrow 5p = 2.84 \Rightarrow p = 0.568$$

$$q = 1 - p = 1 - 0.568 = 0.432$$

**Binomial Distribution Formula:**

$$P(X = x) = {}^{n}C_{x} \cdot p^{x} \cdot q^{n-x} = {}^{5}C_{x} \cdot (0.568)^{x} \cdot (0.432)^{5-x}, \quad x = 0,1,2,\ldots,5$$

$$N = \sum f = 100$$

**Theoretical or Expected Frequency** $f(x) = N \cdot P(X = x)$

$$f(0) = 100 \cdot {}^{5}C_{0} \cdot (0.568)^{0} \cdot (0.432)^{5} = 1.505 \approx 1.5$$

$$f(1) = 100 \cdot {}^{5}C_{1} \cdot (0.568)^{1} \cdot (0.432)^{4} = 9.899 \approx 10$$

$$f(2) = 100 \cdot {}^{5}C_{2} \cdot (0.568)^{2} \cdot (0.432)^{3} = 26.01 \approx 26$$

$$f(3) = 100 \cdot {}^{5}C_{3} \cdot (0.568)^{3} \cdot (0.432)^{2} = 34.2 \approx 34$$

$$f(4) = 100 \cdot {}^{5}C_{4} \cdot (0.568)^{4} \cdot (0.432)^{1} = 22.48 \approx 22$$

$$f(5) = 100 \cdot {}^{5}C_{5} \cdot (0.568)^{5} \cdot (0.432)^{0} = 5.91 \approx 6$$

**Binomial Distribution Table:**

| $x$ | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| **Expected binomial frequency** $f(x)$ | 1.5 | 10 | 26 | 34 | 22 | 6 |

## Poisson Distribution

A random variable $X$ is said to follow Poisson distribution if the probability of $x$ is given by

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

where $\lambda$ is called the **parameter of the distribution**.

## **Poisson Approximation to the Binomial Distribution:**

Poisson distribution is a limiting case of binomial distribution under the following conditions:

(i) The number of trials should be infinitely large, i.e., $n \to \infty$.

(ii) The probability of success $p$ for each trial should be very small, i.e., $p \to 0$.

(iii) $np = \lambda$ should be finite where $\lambda$ is a constant.

The binomial distribution is: $P(X = x) = \binom{n}{x} p^x q^{n-x}$ [since $q$ =$1 - p$]

Putting $p = \frac{\lambda}{n}$,

$$P(X = x) = \frac{n(n-1)(n-2)\ldots(n-x+1)(n-x)!}{(n-x)!\,.\,x!} \left(\frac{\lambda}{n}\right)^x \left(1 - \frac{\lambda}{n}\right)^{n-x}$$

$$P(X = x) = \frac{n(n-1)(n-2)\ldots(n-x+1)}{x!} \left(\frac{\lambda}{n}\right)^x \left(1 - \frac{\lambda}{n}\right)^{n-x}$$

$$= \frac{n(n-1)(n-2)\ldots(n-x+1)}{x!} \frac{\lambda^x}{n^x}\left(1 - \frac{\lambda}{n}\right)^n \left(1 - \frac{\lambda}{n}\right)^{-x}$$

$$= \frac{n(n-1)(n-2)\ldots(n-x+1)}{n^x} \frac{\lambda^x}{x!}\left(1 - \frac{\lambda}{n}\right)^n \left(1 - \frac{\lambda}{n}\right)^{-x}$$

$$= \left[1 \cdot \left(1 - \frac{1}{n}\right) \cdot \left(1 - \frac{2}{n}\right) \ldots \left(1 - \frac{x-1}{n}\right)\right] \frac{\lambda^x}{x!}\left(1 - \frac{\lambda}{n}\right)^n \left(1 - \frac{\lambda}{n}\right)^{-x}$$

Since $lim_{n\to\infty} \left(1 - \frac{\lambda}{n}\right)^{-x} = e^{-\lambda}$ and $lim_{n\to\infty} \left(1 - \frac{1}{n}\right)\left(1 - \frac{2}{n}\right)\ldots\left(1 - \frac{x-1}{n}\right) = 1$

Taking the limits of both the sides as $n \to \infty$,

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

---

## **Examples of Poisson Distribution:**

(i) Number of defective bulbs produced by a reputed company

(ii) Number of telephone calls per minute at a switchboard

(iii) Number of cars passing a certain point in one minute

(iv) Number of printing mistakes per page in a large text

(v) Number of persons born blind per year in a large city

---

**Conditions of Poisson Distribution:**

The Poisson distribution holds under the following conditions:

(i) The random variable $X$ should be discrete.

(ii) The numbers of trials $n$ is very large.

(iii) The probability of success is $p$ is very small (very close to zero).

(iv) $\lambda = np$ is finite.

(v) The occurrences are rare.

---

**Constants of the Poisson Distribution:**

**1. Mean of the Poisson Distribution**

$$E(X) = \sum_{x=0}^{\infty} x\, p(x)$$

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

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

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

$$= \lambda e^{-\lambda}\left(1 + \lambda + \frac{\lambda^2}{2!} + \cdots\right) \quad where\ r = x - 1$$

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

$$= \lambda$$

---

**2. Variance of the Poisson Distribution:**

$$Var(X) = E(X^2) - \mu^2$$

$$= \sum_{x=0}^{\infty} x^2 p(x) - \lambda^2$$

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

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

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

$$= \sum_{x=0}^{\infty} x(x-1) . \frac{e^{-\lambda}\lambda^{x-2}\lambda^2}{x(x-1)(x-2)! \dots} + \lambda - \lambda^2 \quad [Since \sum_{x=0}^{\infty} x \cdot \frac{e^{-\lambda}\lambda^x}{x!} = \lambda]$$

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

$$= e^{-\lambda}\lambda^2 \left( 1 + \lambda + \frac{\lambda^2}{2!} + \cdots \right) + \lambda - \lambda^2$$

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

$$= \lambda^2 + \lambda - \lambda^2$$

$$= \lambda$$

---

**3. Standard Deviation of the Poisson Distribution:**

$$SD = \sqrt{Variance} = \sqrt{\lambda}$$

---

**4. Mode of the Poisson Distribution:**

Mode is the value of $x$ for which the probability $p(x)$ is maximum.

$$p(x) \geq p(x+1) \quad and \quad p(x) \geq p(x-1)$$

When $p(x) \geq p(x+1)$,

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

$$1 \geq \frac{\lambda}{x+1}$$

$$(x+1) \geq \lambda$$

$$x \geq \lambda - 1 \ldots (1)$$

Similarly, for $p(x) \geq p(x-1)$

$$x \leq \lambda \ldots . (2)$$

Comparing eq(1) and (2) we get $\lambda - 1 \leq x \leq \lambda$

Hence, the mode of the Poisson distribution lies between $\lambda - 1$ and $\lambda$.

**Case I**: If $\lambda$ is an integer then $\lambda - 1$ is also an integer. The distribution is bimodal and the two modes are $\lambda - 1$ and $\lambda$.

**Case II**: If $\lambda$ is not an integer, the distribution is unimodal and the mode of the Poisson distribution is an integral part of $\lambda$. The mode is the integer between $\lambda - 1$ and $\lambda$.

---

**Recurrence Relation for the Poisson Distribution:**

For the Poisson distribution,

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

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

$$= \frac{\lambda}{x+1} \cdot \frac{e^{-\lambda}\lambda^{x}}{x!}$$

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

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

---

**Problem1:** If the mean of a Poisson variable is 1.8, find (i) P(X > 1), (ii) P(X = 5), and

(iii) P(0 < X < 5).

**Solution** For a Poisson distribution,

$$\lambda = 1.8$$

$$P(X=x)=\frac{e^{-1.8}(1.8)^x}{x!}, \quad x=0,1,2,\dots$$

(i)

$$P(X>1)=1-P(X\leq 1)$$

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

$$=1-\sum_{x=0}^{1} \frac{e^{-1.8}(1.8)^x}{x!}$$

$$=1-\left(e^{-1.8}+\frac{1.8e^{-1.8}}{1!}\right)$$

$$=0.5372$$

(ii)

$$P(X=5)=\frac{e^{-1.8}(1.8)^5}{5!}=0.026$$

(iii)

$$P(0<X<5)=P(X=1)+P(X=2)+P(X=3)+P(X=4)$$

$$=\sum_{x=1}^{4} \frac{e^{-1.8}(1.8)^x}{x!}$$

$$=0.7983$$

---

**Problem2:** If a random variable has a Poisson distribution such that P(X = 1) = P(X = 2), find (i) the mean of the distribution, (ii) P(X = 4), (iii) P(X ≥ 1), and (iv) P(1 < X < 4).

**Solution** For a Poisson distribution,

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

(i)

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

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

$$\lambda = 2$$

Hence,

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

(ii)

$$P(X = 4) = \frac{e^{-2}2^4}{4!}$$

$$= 0.9022$$

(iii)

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

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

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

$$= 0.8647$$

(iv)

$$P(1 < X < 4) = P(X = 2) + P(X = 3)$$

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

$$= 0.4511$$

---

**Problem3:** If X is a Poisson variate such that P(X = 0) = P(X = 1), find P(X = 0) and using recurrence relation formula, find the probabilities at x = 1, 2, 3, 4, and 5.

**Solution** For a Poisson distribution,

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

$$P(X = 0) = P(X = 1)$$

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

$$1 = \lambda$$

Hence,

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

(i)

$$P(X = 0) = \frac{e^{-1}1^0}{0!} = 0.3678$$

(ii) By recurrence relation,

$$p(x+1) = \frac{\lambda}{x+1}p(x) \quad [\lambda = 1]$$

$$p(1) = p(0) = 0.3678$$

$$p(2) = \frac{1}{2}(0.3678) = 0.1839$$

$$p(3) = \frac{1}{3}(0.1839) = 0.0613$$

$$p(4) = \frac{1}{4}(0.0613) = 0.015325$$

$$p(5) = \frac{1}{5}(0.015325) = 0.003065$$

---

**Problem4:** If the variance of a Poisson variate is 3, find the probability that (i) X = 0, (ii) $0 < X \leq 3$, and (iii) $1 \leq X < 4$.

**Solution** For a Poisson distribution,

$$Variance = Mean = \lambda = 3$$

$$P(X = x) = \frac{e^{-3}3^x}{x!}, \quad x = 0,1,2, \ldots$$

(i)

$$P(X = 0) = \frac{e^{-3}3^0}{0!} = e^{-3} = 0.0498$$

(ii)

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

$$= \sum_{x=1}^{3} \frac{e^{-3}3^x}{x!}$$

$$= 0.5974$$

(iii)

$$P(1 \le X < 4) = P(X = 1) + P(X = 2) + P(X = 3)$$

$$= \sum_{x=1}^{3} \frac{e^{-3}3^x}{x!}$$

$$= 0.5974$$

---

**Problem5:** If a Poisson distribution is such that

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

find (i) $P(X \ge 1)$, (ii) $P(X \le 3)$, and (iii) $P(2 \le X \le 5)$.

**Solution** For a Poisson distribution,

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

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

$$\frac{3}{2} \cdot \frac{e^{-\lambda}\lambda}{1!} = \frac{e^{-\lambda}\lambda^3}{3!}$$

$$\frac{3}{2} = \frac{\lambda^2}{6}$$

$$\lambda^2 - 9 = 0$$

$$\lambda = 3, -3$$

Since $\lambda > 0$, $\lambda = 3$.

Hence,

$$P(X = x) = \frac{e^{-3}3^x}{x!}, \quad x = 0,1,2, \ldots$$

(i)

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

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

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

$$= 0.9502$$

(ii)

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

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

$$= 0.6472$$

(iii)

$$P(2 \leq X \leq 5) = P(X = 2) + P(X = 3) + P(X = 4) + P(X = 5)$$

$$= \sum_{x=2}^{5} \frac{e^{-3}3^x}{x!}$$

$$= 0.7169$$

---

**Problem6:** Six coins are tossed 6400 times. Using the Poisson distribution, what is the approximate probability of getting six heads 10 times?

(ii) A Poisson distribution has a double mode at $x = 3$ and $x = 4$. What is the probability

that $x$ will have one or the other of· these two values.

**Solution**

Let $p$ be the probability of getting one head with one coin.

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

Probability of getting 6 heads with 6 coins = $\left(\frac{1}{2}\right)^6 = \frac{1}{64}$

$$n = 6400$$

$$\lambda = np = 6400\left(\frac{1}{64}\right) = 100$$

Probability of getting $x$ heads

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

Probability of getting 6 heads 10 times

$$P(X = 10) = \frac{e^{-100}(100)^{10}}{10!} = 1.025 \times 10^{-30}$$

(*ii*) Since 2 modes are given when $\lambda$ is an integer, modes are $\lambda - 1$ and $\lambda$.

$$\lambda - 1 = 3 \Rightarrow \lambda = 4$$

Probability (when $x = 3$) $= \frac{e^{-4}(4)^3}{3!}$

Probability (when $x = 4$) $= \frac{e^{-4}(4)^4}{4!}$

Required Probability $= P(x = 3 \text{ } or \text{ } 4) = P(x = 3) + P(x = 4)$

$$= \frac{e^{-4}(4)^3}{3!} + \frac{e^{-4}(4)^4}{4!} = \frac{64}{3}e^{-4} = 0.39073.$$

---

**Problem7:** If 2% of lightbulbs are defective, find the probability that (i) at least one is defective, and (ii) exactly 7 are defective. Also, find $P(1 < X < 8)$ in a sample of 100.

**Solution**

Let $p$ be the probability of defective bulb.

$$p = 2\% = 0.02$$

$$n = 100$$

Since $p$ is very small and $n$ is large, Poisson distribution is used.

$$\lambda = np = 100(0.02) = 2$$

Let $X$ be the random variable which denotes the number of defective bulbs in a sample of 100.

Probability of $x$ defective bulbs in a sample of 100

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

(i) Probability that at least one bulb is defective

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

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

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

$$= 0.8647$$

---

(ii) Probability that exactly 7 bulbs are defective

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

$$= 0.0034$$

---

(iii) $P(1 < X < 8) = P(X = 2) + P(X = 3) + P(X = 4) + P(X = 5) + P(X = 6) + P(X = 7)$

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

$$= 0.5929$$

---

**Problem8:** A hospital switchboard receives an average of 4 emergency calls in a 10-minute interval. What is the probability that (i) there are at most 2 emergency calls? (ii) there are exactly 3 emergency calls in an interval of 10 minutes?

**Solution**

Let $p$ be the probability of receiving emergency calls per minute.

$$p = \frac{4}{10} = 0.4, \quad n = 10$$

$$\lambda = np = 10(0.4) = 4$$

Let $X$ be the random variable which denotes the number of emergency calls per minute.

Probability of $x$ emergency calls per minute:

$$P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!} = \frac{e^{-4}4^x}{x!}, \quad x = 0,1,2, \dots$$

---

(i) Probability that there are at most 2 emergency calls

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

$$= \sum_{x=0}^{2} \frac{e^{-4}4^x}{x!}$$

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

$$= 0.238$$

---

(ii) Probability that there are exactly 3 emergency calls

$$P(X = 3) = \frac{e^{-4}4^3}{3!}$$

$$= 0.1954$$

---

**Problem8:** The number of accidents in a year attributed to taxi drivers in a city follows Poisson distribution with a mean of 3. Out of 1000 taxi drivers, find approximately the number of drivers with (i) no accidents in a year, and (ii) more than 3 accidents in a year.

**Solution**

For a Poisson distribution,

$$\lambda = 3, \quad N = 1000$$

Probability of $x$ accidents in a year:

$$P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!} = \frac{e^{-3}3^x}{x!}, \quad x = 0,1,2, \dots$$

---

(i) Probability of no accidents in a year:

$$P(X = 0) = \frac{e^{-3}3^0}{0!} = e^{-3}$$

$$= 0.0498$$

Number of drivers with no accidents:

$$f(x) = NP(X = 0) = 1000(0.0498) = 49.8 \approx 50$$

---

(ii) Probability of more than 3 accidents in a year:

$$P(X > 3) = 1 - P(X \le 3)$$

$$= 1 - [P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3)]$$

$$= 1 - \sum_{x=0}^{3} P(X = x)$$

$$= 1 - \sum_{x=0}^{3} \frac{e^{-3}3^x}{x!}$$

$$= 0.3528$$

Number of drivers with more than 3 accidents:

$$f(x) = NP(X > 3) = 1000(0.3528) = 352.8 \approx 353$$

---

**Problem9: fit a Poisson distribution** to the given data:

| Number of deaths (x) | 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| Frequency (f) | 122 | 60 | 15 | 2 | 1 |

---

Step 1: Calculate the Mean (λ)

The mean λ of a Poisson distribution is:

$$\lambda = \frac{\sum f x}{\sum f}$$

$$\lambda = \frac{122(0) + 60(1) + 15(2) + 2(3) + 1(4)}{122 + 60 + 15 + 2 + 1}$$

$$\lambda = \frac{0 + 60 + 30 + 6 + 4}{200} = \frac{100}{200} = 0.5$$

So, **λ = 0.5**

---

Step 2: Poisson Probability Formula

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

For λ = 0.5:

- **P(0)** = $e^{-0.5} \cdot \frac{0.5^0}{0!} = e^{-0.5} = 0.6065$
- **P(1)** = $e^{-0.5} \cdot \frac{0.5^1}{1!} = 0.6065 \cdot 0.5 = 0.3033$
- **P(2)** = $e^{-0.5} \cdot \frac{0.5^2}{2!} = 0.6065 \cdot 0.125 = 0.0758$
- **P(3)** = $e^{-0.5} \cdot \frac{0.5^3}{3!} = 0.6065 \cdot 0.0208 = 0.0126$
- **P(4)** = $e^{-0.5} \cdot \frac{0.5^4}{4!} = 0.6065 \cdot 0.0026 = 0.0016$

---

Step 3: Expected Frequencies

Expected frequency = $N \cdot P(X = x)$, where $N = \sum f = 200$.

- **E(0) = 200 × 0.6065 = 121.3 ≈ 121**
- **E(1) = 200 × 0.3033 = 60.7 ≈ 61**
- **E(2) = 200 × 0.0758 = 15.2 ≈ 15**
- **E(3) = 200 × 0.0126 = 2.5 ≈ 3**
- **E(4) = 200 × 0.0016 = 0.3 ≈ 0**

---

Step 4: Comparison Table

| x | Observed f | Poisson P(x) | Expected f (N·P(x)) |
|---|---|---|---|
| 0 | 122 | 0.6065 | 121 |
| 1 | 60 | 0.3033 | 61 |
| 2 | 15 | 0.0758 | 15 |
| 3 | 2 | 0.0126 | 3 |
| 4 | 1 | 0.0016 | 0 |

---

## NORMAL DISTRIBUTION

The normal distribution is a continuous distribution. It can be derived from the binomial distribution in the limiting case when $n$, the number of trials is very large and $p$, the probability of a success, is close to $\frac{1}{2}$. The general equation of the normal distribution is given by

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

where the variable $x$ can assume all values from $-\infty$ $to + \infty$. $\mu$ and $\sigma$, called the parameters of the distribution, are respectively the mean and the standard deviation of the distribution and $-\infty < \mu < \infty, \sigma > 0$. *Here* $x$ is called the normal variate and $f(x)$ is called probability density function of the normal distribution.

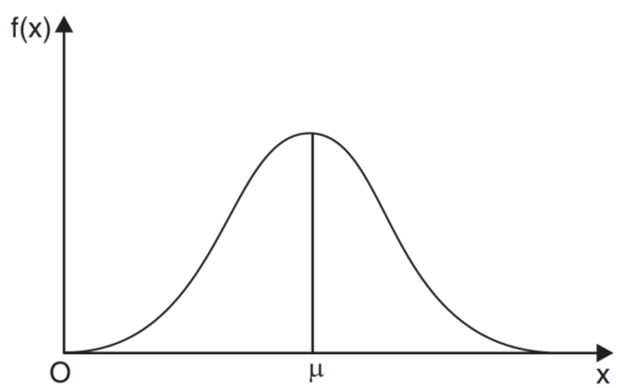
If a variable $x$ has the normal distribution with mean $\mu$ and standard deviation $\sigma$, we

briefly write $x : N(\mu, \sigma^2)$

The graph of the normal distribution is called the *normal curve*. It is bell-shaped and

symmetrical about the mean $\mu$. The two tails of the curve extend $to + \infty$ $and - \infty$ towards the positive and negative directions of the $x$-axis respectively and gradually approach the $x$-axis without ever meeting it. The curve is unimodal and the mode of the normal distribution coincides with its mean $\mu$. The line $x = \mu$ divides the area under the normal curve above $x$-axis into two equal parts. Thus, the median of the distribution also coincides with its mean and mode. The area under the normal curve between any two given ordinates $x = x_1$ and $x = x_2$

represents the probability of values falling into the given interval. The total area under the normal curve above the $x$ $-$axis is 1.




---

**BASIC PROPERTIES OF THE NORMAL DISTRIBUTION:**

The probability density function of the normal distribution is given by

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

$(i)$ $f(x) \geq 0$ $(ii)$ $\int_{-\infty}^{\infty} f(x)dx = 1$,

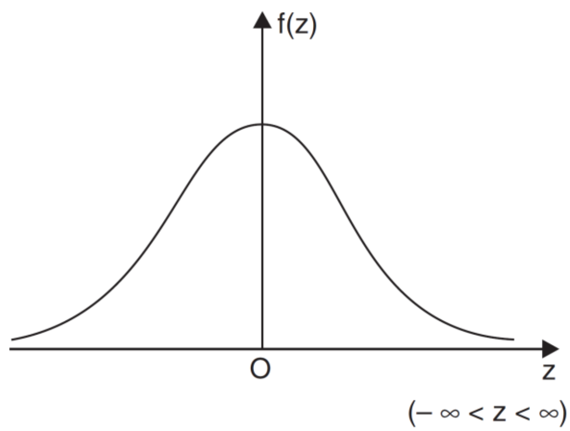
*i. e.*, the total area under the normal curve above the $x$ −axis is 1.

$(iii)$ The normal distribution is symmetrical about its mean.

$(iv)$ It is a unimodal distribution. The mean, mode and median of this distribution coincide.

---

**STANDARD FORM OF THE NORMAL DISTRIBUTION:**




If X is a normal random variable with mean $\mu$ and standard deviation $\sigma$, then the random variable $Z = \frac{X-\mu}{\sigma}$ has the normal distribution with mean 0 and standard deviation 1. The random variable $Z$ is called the *standardized* (or *standard*) normal random variable.

The probability density function for the normal distribution in standard form is given by

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}$$

It is free from any parameter. This helps us to compute areas under the normal probability curve by making use of standard tables.

**Note 1.** If $f(z)$ is the probability density function for the normal distribution, then

$P(z_1 \leq Z \leq z_2) = \int_{z_1}^{z_2} f(z)dz = F(z_2) - F(z_1)$ , where $F(z) = \int_{-\infty}^{z} f(z)dz = P(Z \leq z)$

The function $F(z)$ defined above is called the *distribution function* for the normal distribution.

**Note 2.** The probabilities $P(z_1 \le Z \le z_2)$, $P(z_1 < Z \le z_2)$, $P(z_1 \le Z < z_2)$ and

$P(z_1 < Z < z_2)$ are all regarded to be the same.

**Note 3.** $F(-z_1) = 1 - F(z_1)$

---

**Characteristics of normal distribution:**

1. The normal distribution curve is bell shaped curve and it is symmetric about the line $x = \mu$
2. X – axis is an asymptote to the normal curve.
3. Area under normal curve is unity.
4. The linear combination of independent normal variate is a normal variate. i.e if x and y are normal variate then ax + by is also a normal variate.
5. No portion of the curve lies below the x – axis.
6. The mean, median and mode of the N.D. will coincide.
7. $p\ (\mu - \sigma \le x \le \mu + \sigma) = p\ (-1 \le x \le 1) = 0.6826$
8. $p\ (\mu - 2\sigma \le x \le \mu + 2\sigma) = p\ (-2 \le x \le 2) = 0.9844$
9. $p\ (\mu - 3\sigma \le x \le \mu + 3\sigma) = p\ (-3 \le x \le 3) = 0.9973$

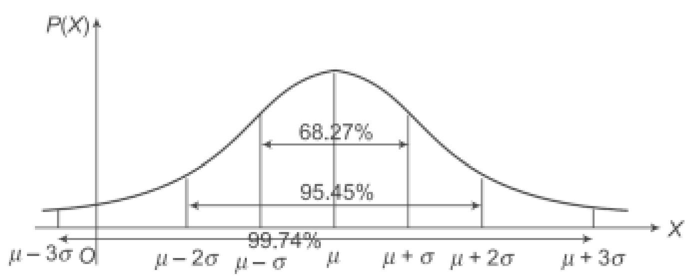



Fig. 5.2

10. The max. value of Normal density (also called Gaussian distribution) is $\frac{1}{\sigma\sqrt{2\pi}}$ and it will occur at $x = \mu$

**Proof:** we know that $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$

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

$$f'(x) = f(x)\left[-\left(\frac{x-\mu}{\sigma^2}\right)\right] \Rightarrow 0 = f(x)[-\left(\frac{x-\mu}{\sigma^2}\right)] \Rightarrow -\left(\frac{x-\mu}{\sigma^2}\right) = 0$$

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

At $x = \mu$ we will get $f$ maximum

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

**Note1:** $\Gamma n = \int_0^\infty e^{-x} x^{n-1} dx$

**Note2:** $\Gamma\frac{1}{2} = \sqrt{\pi}$

**Note3:** $\Gamma n+1 = n\,\Gamma n$

**Note4:** $\int_0^\infty e^{-z^2/2} dz = \sqrt{\frac{\pi}{2}}$

**Proof:** Put $\frac{z^2}{2} = t \Rightarrow z^2 = 2t \Rightarrow z = \sqrt{2t} \Rightarrow dz = \sqrt{2}\frac{1}{2\sqrt{t}} dt = \frac{1}{\sqrt{2t}} dt$

Substitute in the above we have $\int_0^\infty e^{-t} \frac{1}{\sqrt{2t}} dt = \frac{1}{\sqrt{2}} \int_0^\infty e^{-t} t^{-\frac{1}{2}} dt =$
$\frac{1}{\sqrt{2}} \int_0^\infty e^{-t} t^{\frac{1}{2}-1} dt = \frac{1}{\sqrt{2}} \Gamma\frac{1}{2} = \frac{1}{\sqrt{2}} \sqrt{\pi} = \sqrt{\frac{\pi}{2}}$

---

**MEAN OF NORMAL DISTRIBUTION**

Consider the normal distribution with $\mu$, $\sigma$ as the parameters then

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

$= \int_{-\infty}^{\infty} x.\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} dx$ [ Since $\int_{-\infty}^{\infty} f(x)dx$ =area under normal curve $= 1$]

Put $\frac{x-\mu}{\sigma} = z$ so that $x = \mu + \sigma z$ $\therefore dx = \sigma dz$

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

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

$$\frac{2\mu}{\sqrt{2\pi}} \int_0^\infty e^{-z^2/2} dz = \frac{2\mu}{\sqrt{2\pi}} \sqrt{\frac{\pi}{2}} = \mu$$

[ since $\int_{-\infty}^{\infty} z e^{-\frac{1}{2}z^2} dz$ is odd function $\Rightarrow \int_{-\infty}^{\infty} z e^{-\frac{1}{2}z^2} dz = 0$ and $\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz$ is even function so, we write $\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz = \frac{2}{\sqrt{2\pi}} \int_0^\infty e^{-z^2/2} dz$ and also we know that $\int_0^\infty e^{-z^2/2} dz = \sqrt{\frac{\pi}{2}}$ ]

## VARIANCE OF NORMAL DISTRIBUTION

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

Put $\frac{x-\mu}{\sigma} = z \;\Rightarrow x - \mu = \sigma z \Rightarrow \; x = \mu + \sigma z \qquad \therefore dx = \sigma dz$

Variance = $\int_{-\infty}^{\infty} \quad (\sigma z)^2 \frac{1}{\sigma\sqrt{2\pi}} e^{-z^2/2} \sigma dz$

$$= \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \quad (z)^2 e^{-z^2/2} \sigma dz$$

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

Put $\frac{z^2}{2} = t \Rightarrow z^2 = 2t \Rightarrow z = \sqrt{2t} \Rightarrow dz = \sqrt{2}\frac{1}{2\sqrt{t}} dt = \frac{1}{\sqrt{2t}} dt$

$$= \frac{2\sigma^2}{\sqrt{2\pi}} \int_{0}^{\infty} \quad 2te^{-t} \frac{dt}{\sqrt{2t}}$$

$$= \frac{2\sigma^2}{\sqrt{\pi}} \int_{0}^{\infty} \quad te^{-t} \frac{dt}{\sqrt{t}}$$

$$= \frac{2\sigma^2}{\sqrt{\pi}} \int_{0}^{\infty} \quad \sqrt{t} e^{-t} dt$$

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

$$\frac{2\sigma^2}{\sqrt{\pi}} \Gamma\left(\frac{3}{2}\right) = \frac{2\sigma^2}{\sqrt{\pi}} \frac{1}{2} \Gamma\left(\frac{1}{2}\right) = \frac{\sigma^2}{\sqrt{\pi}} \sqrt{\pi} = \sigma^2$$

---

**The standard deviation of the normal distribution is** $\sigma$

---

**Mode of normal distribution:**

Mode is the value of x for which $f(x)$ maximum i.e at mode $f'(x) = 0$ and $f''(x) < 0$

By definition we have $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$

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

$$f'(x) = f(x)\left[-\left(\frac{x-\mu}{\sigma^2}\right)\right] = [-\frac{1}{\sigma^2} f(x)(x-\mu)]$$

Now, $f'(x) = 0$ i.e $x = \mu$

$$f''(x) = [f'(x)'] = [-\frac{1}{\sigma^2} f(x)(x-\mu)]' = -\frac{1}{\sigma^2}[f(x)(x-\mu)]'$$

$$= -\frac{1}{\sigma^2}[(x-\mu)f'(x) + f(x)]$$

$$= -\frac{1}{\sigma^2}[(x-\mu)(-f(x))\left[\left(\frac{x-\mu}{\sigma^2}\right)\right] + f(x)]$$

$$= -\frac{f(x)}{\sigma^2}[1 - \left[\left(\frac{(x-\mu)^2}{\sigma^2}\right)\right]]$$

At $x = \mu$, $f''(x)$ becomes negative i.e at $x = \mu$ $f''(x) = -\frac{f(x)}{\sigma^2} < 0$

Hence $x = \mu$ is mode of normal distribution

---

**Median of normal distribution:**

If M median of normal distribution $\int_{-\infty}^{M} f(x) = \frac{1}{2}$

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

i.e $\frac{1}{\sigma\sqrt{2\pi}} [\int_{-\infty}^{\mu} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} dx + \int_{\mu}^{M} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} dx] = \frac{1}{2}$.........(1)

Consider, $\frac{1}{\sigma\sqrt{2\pi}} [\int_{-\infty}^{\mu} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} dx]$

Put $\frac{x-\mu}{\sigma} = z$ so that $x = \mu + \sigma z$ $\quad \therefore dx = \sigma dz$

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

$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{0} e^{-\frac{z^2}{2}} dz$ also $\frac{1}{\sqrt{2\pi}} \int_{0}^{\infty} e^{-\frac{z^2}{2}} dz$ ( by symmetry)

$= \frac{1}{\sqrt{2\pi}} \sqrt{\frac{\pi}{2}} = \frac{1}{2}$.....(2)

Sub (2) in (1) we get

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

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

$\Rightarrow x=\mu$ [ since $\int_{\mu}^{M} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}dx$ becomes 0 if either $x=\mu$ or integrand = 0 or limits are equal]

Therefore, in N.D Mean = Median = Mode

---

**Mean deviation from the mean for normal distribution:**

By the definition of Mean deviation, we have $\int_{-\infty}^{\infty} |(x-\mu)| f(x)dx$

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

[ put Put $\frac{x-\mu}{\sigma}=z$ so that $x=\mu+\sigma z \quad \therefore dx=\sigma dz$]

$$\frac{\sigma}{\sqrt{2\pi}}\int_{-\infty}^{\infty} |z| e^{-\frac{1}{2}(z)^2}dz$$

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

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

[put Put $\frac{z^2}{2}=t \Rightarrow z^2=2t \Rightarrow 2zdz=2dt \Rightarrow zdz=dt$]

$$=\frac{\sqrt{2}\,\sigma}{\sqrt{\pi}}\int_{0}^{\infty} e^{-t}dt=\frac{\sqrt{2}\,\sigma}{\sqrt{\pi}}\left(\frac{e^{-t}}{-1}\right)_0^{\infty}$$

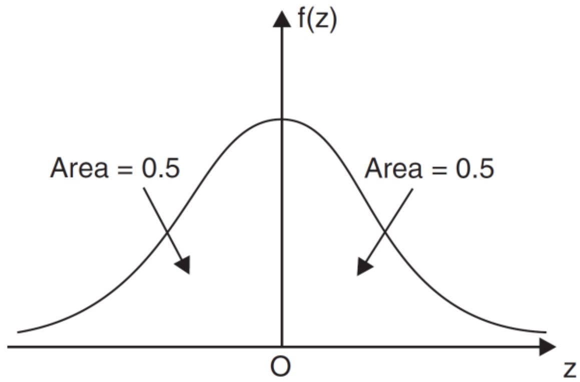
$\frac{\sqrt{2}\,\sigma}{\sqrt{\pi}}(0+1)=\frac{\sqrt{2}\,\sigma}{\sqrt{\pi}}=\frac{4}{5}\sigma$ (approximetly)

---

**AREA UNDER THE NORMAL CURVE:**

$The\ total$ area und$er\ this$ curve is $1\ By\ t$aking $z = \frac{x-\mu}{\sigma}$, standard n$ormal$ curve is formed.

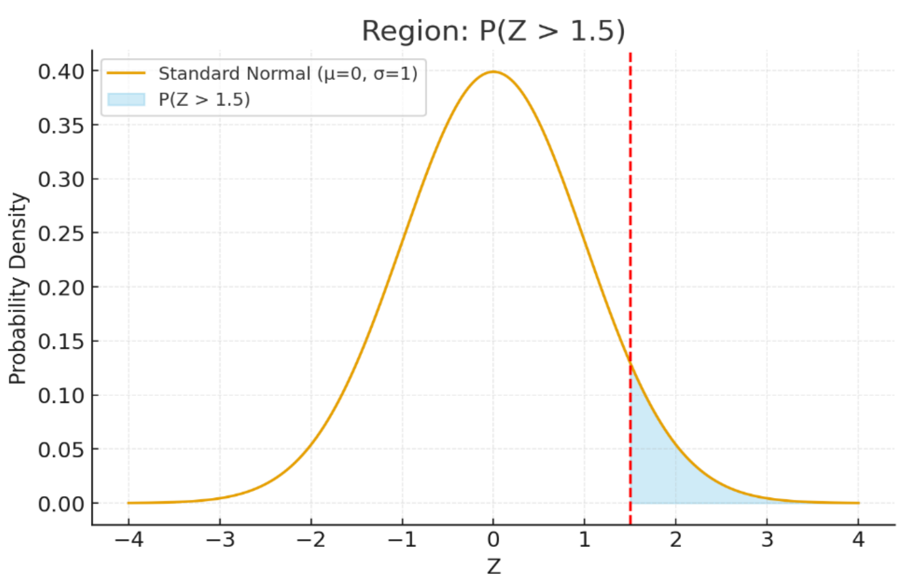
The area under the curve is divided into two equal parts by $z = 0$. The area between the ordinate $z = 0$ and any other ordinate can be noted from the supplied table. It should be noted that mean for the normal distribution is 0.




**APPLICATIONS OF NORMAL DISTRIBUTION:**

De Moivre made the discovery of this distribution in 1733. This distribution has an important application in the theory of errors made by chance in experimental measurements. Its more applications are in computation of hit probability of a shot and in statistical inference in almost every branch of science.

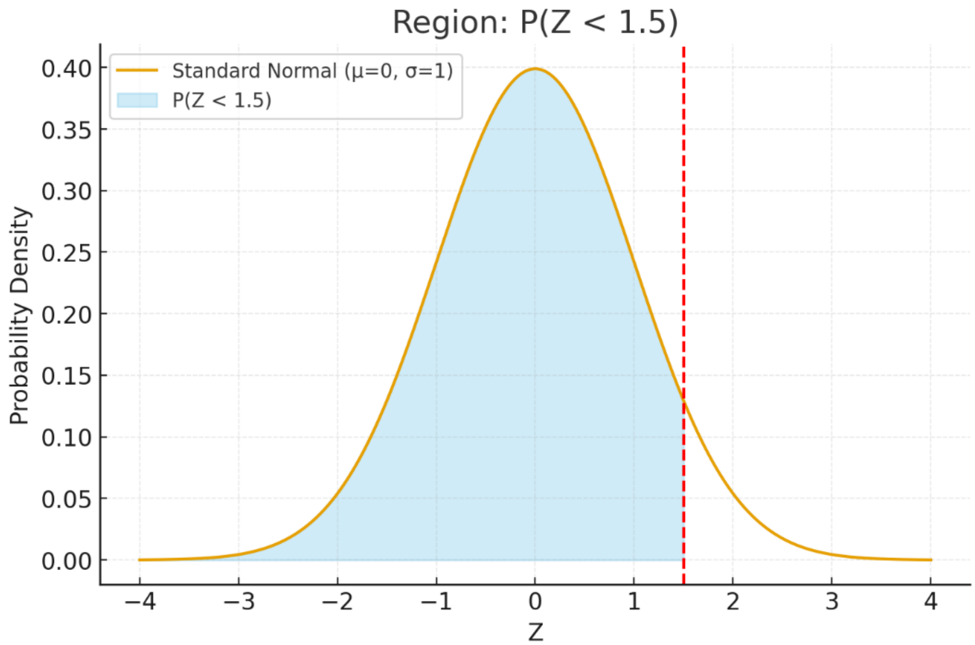
**Method1: When** $z > positive\ values$




$$P(Z > 1.5) \rightarrow \text{shaded area to the right of } Z = 1.5$$

$$p(z > 1.5) = 0.5 - p(0 < z < 1.5) = 0.5 - 0.4332 = 0.0668$$

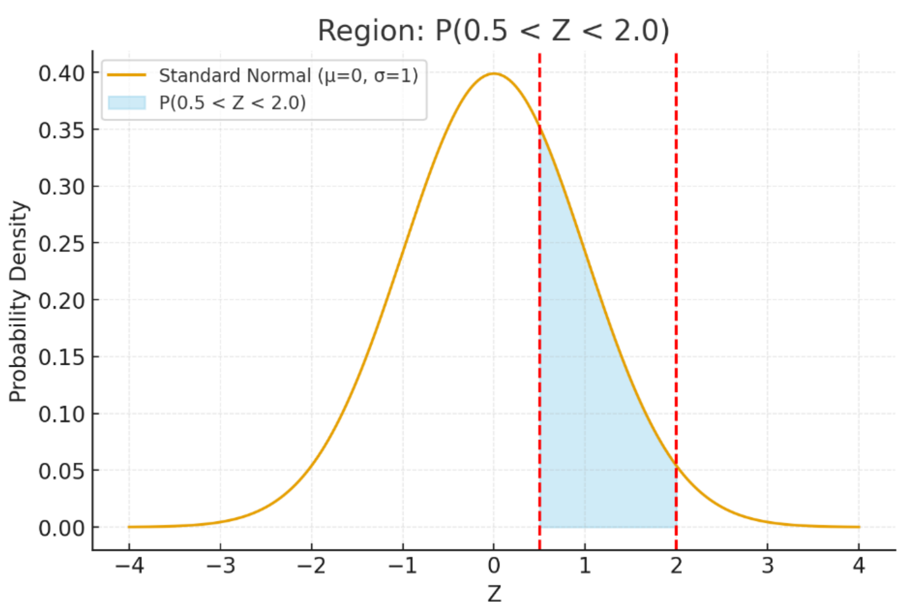
**Method2: When** $z < positive\ values$




$P(Z < 1.5) \rightarrow$ shaded area to the left of $Z = 1.5$.

$$p(z < 1.5) = 0.5 + p(0 < z < 1.5) = 0.5 + 0.4332 = 0.9332$$

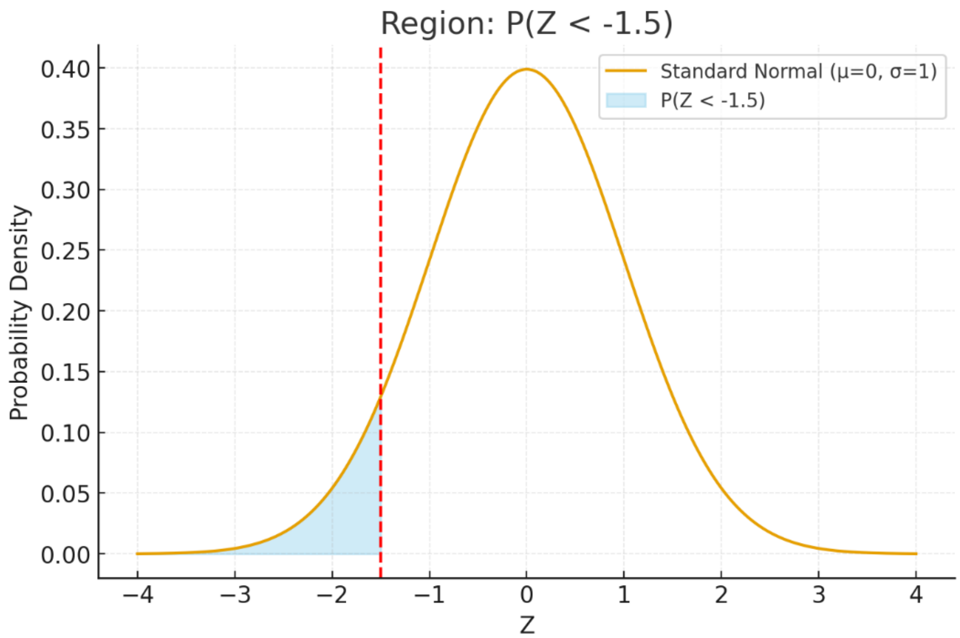
**Method3: When** $z\ lies\ between\ two\ positive\ values$




$P(0.5 < Z < 2.0) \rightarrow$ shaded area between $Z = 0.5$ and $Z = 2.0$

$$P(0.5 < Z < 2.0) = P(0 < Z < 2) - \ p(0 < z < 0.5) = 0.4772 - 0.1916 = 0.285$$

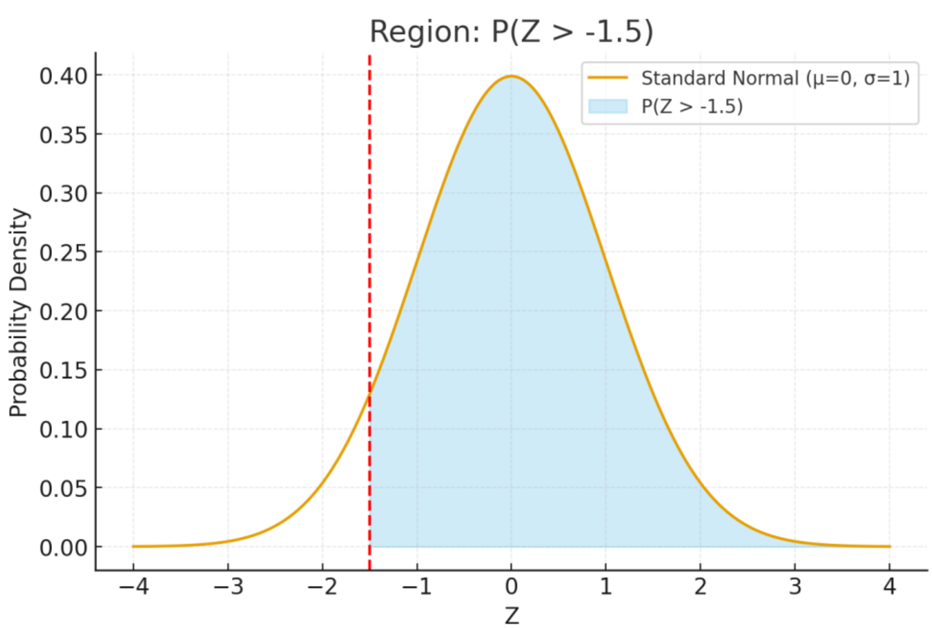
**Method4: When** $z < negative\ values$




$P(Z < -1.5) \rightarrow$ shaded area to the **left** of $-1.5$

$$p(z < -1.5) = 0.5 - p(-1.5 < z < 0) = 0.5 - p(0 < z < 1.5) = 0.5 - 0.4332 = 0.0668$$

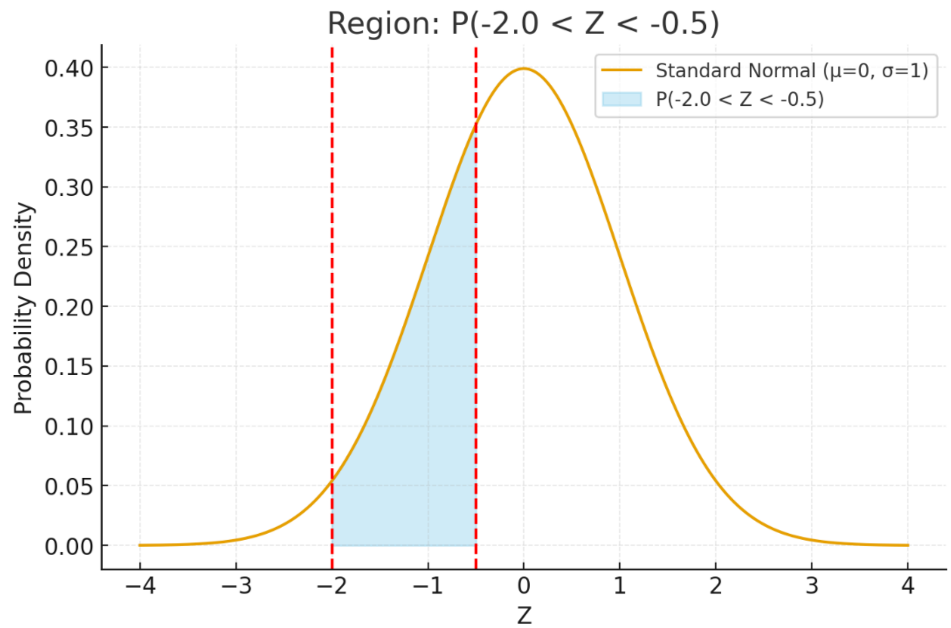
**Method5: When** $z > negative\ values$




$P(Z > -1.5) \rightarrow$ shaded area to the **right** of $-1.5$

$$p(z > -1.5) = 0.5 + p(-1.5 < z < 0) = 0.5 + p(0 < z < 1.5) = 0.5 - 0.4332 = 0.0668$$

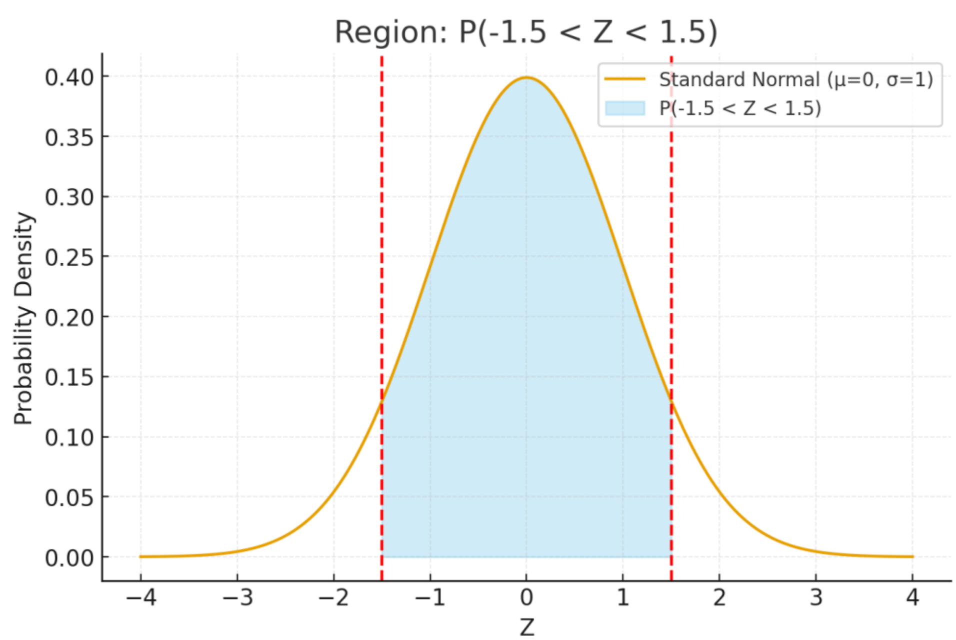
**Method6: When** $z$ *lies between two negative values*




$P(-2.0 < Z < -0.5) \rightarrow$ shaded area between two **negative Z values**

$P(-2.0 < Z < -0.5) = P(0 < Z < 2) - \ p(0 < z < 0.5) = 0.4772 - 0.1916 = 0.2856$

**Method7: When** $z$ *lies betweemt both positive and negative values*




$P(-1.5 < Z < 1.5) \rightarrow$ shaded area between **symmetric positive and negative values**

$$P(-1.5 < Z < 1.5) = P(0 < Z < 1.5) + \; p(0 < z < 1.5) = 0.4332 + 0.4332 = 0.8664$$

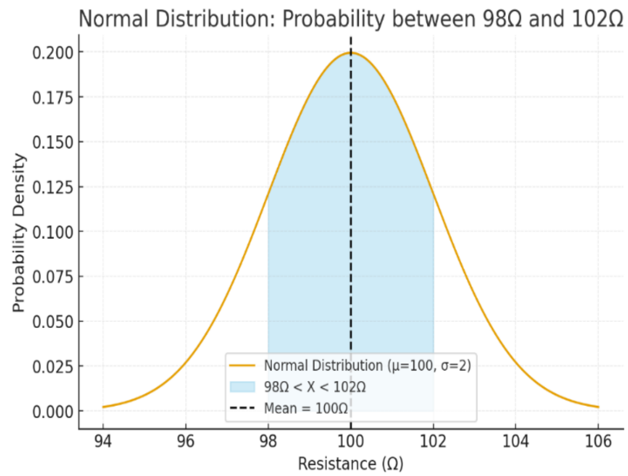
---

**1Q. A manufacturer knows from experience that the resistance of resistors he produces is normal with mean 100 ohms and standard deviation 2 ohms. What percentage of resistors will have resistance between 98 ohms and 102 ohms?**

---

**Solution:** Step 1: Write given data

- Mean ($\mu$) = 100
- Standard deviation ($\sigma$) = 2
- Range = [98, 102]




Step 1: Convert to Z-scores

For $X = 98$:

$$Z_1 = \frac{98 - 100}{2} = \frac{-2}{2} = -1$$

For $X = 102$:

$$Z_2 = \frac{102 - 100}{2} = \frac{2}{2} = 1$$

So,

$$P(98 < X < 102) = P(-1 < Z < 1)$$

---

Step 2: Break into two parts

$$P(-1 < Z < 1) = P(-1 < Z < 0) + P(0 < Z < 1)$$

---

Step 3: Use standard normal values

From Z-tables:

- $P(0 < Z < 1) = 0.3413$
- So $P(-1 < Z < 0) = 0.3413$ (by symmetry)

Thus,

$$P(-1 < Z < 1) = 0.3413 + 0.3413 = 0.6826$$

---

Step 4: Convert probability to percentage

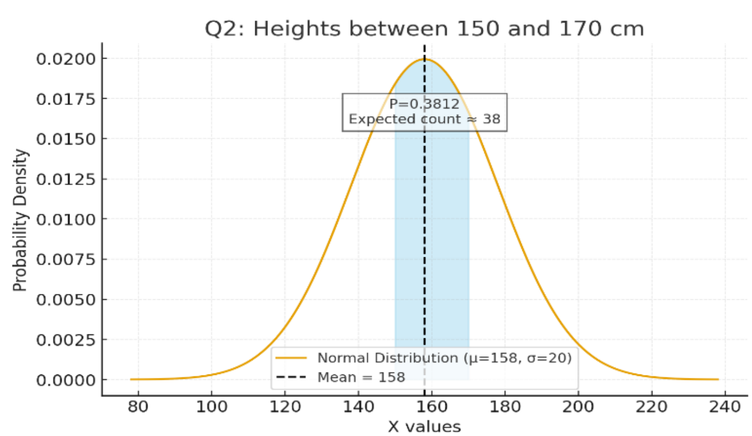
$$0.6826 \times 100 = 68.26\%$$

---

**Q2. Mean height of students in a class is 158 cm with standard deviation of 20 cm. Find how many students' heights lie between 150 cm and 170 cm, if there are 100 students.**

---

**Solution:** Write given data; Mean ($\mu$) = 158; Standard deviation ($\sigma$) = 20; Range = [150, 170]; Total students = 100




Step 1: Convert to Z-scores

For $X = 150$:

$$Z_1 = \frac{150 - 158}{20} = \frac{-8}{20} = -0.4$$

For $X = 170$:

$$Z_2 = \frac{170 - 158}{20} = \frac{12}{20} = 0.6$$

So,

$$P(150 < X < 170) = P(-0.4 < Z < 0.6)$$

---

Step 2: Split into two parts

$$P(-0.4 < Z < 0.6) = P(-0.4 < Z < 0) + P(0 < Z < 0.6)$$

---

Step 3: Use standard normal table values

From Z-table:

- $P(0 < Z < 0.4) = 0.1554 \;\; \Rightarrow \;\; P(-0.4 < Z < 0) = 0.1554$
- $P(0 < Z < 0.6) = 0.2257$

So,

$$P(-0.4 < Z < 0.6) = 0.1554 + 0.2257 = 0.3811$$

---

Step 4: Convert probability to number of students

$$100 \times 0.3811 = 38.1 \approx 38$$

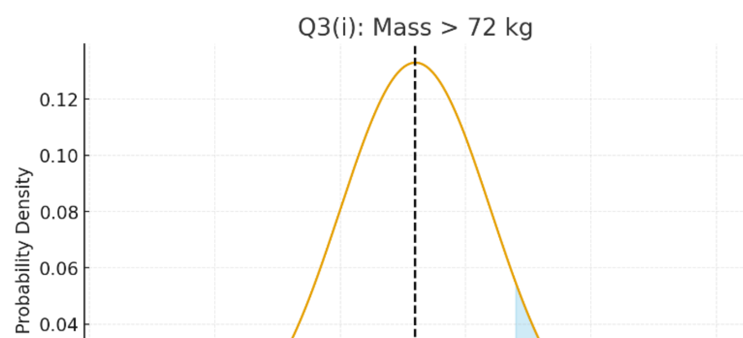
---

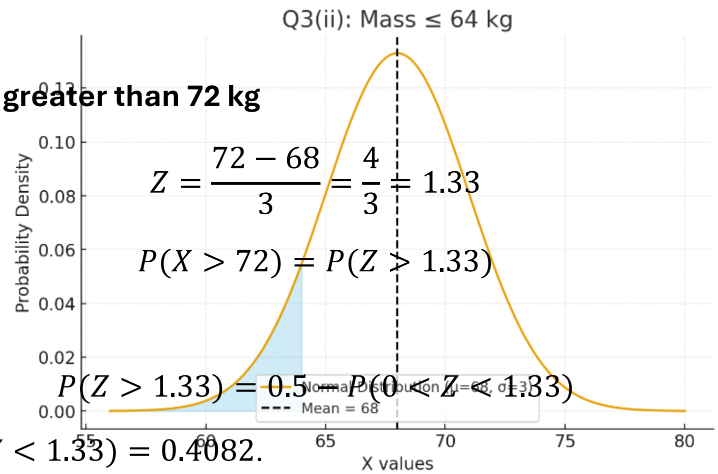
**Q3**. If the masses of 300 students are normally distributed with mean 68 kg and S.D. 3 kg, how many students have masses

- greater than 72 kg
- less than or equal to 64 kg
- between 65 kg and 71 kg inclusive

---

**Solution:** If masses of 300 students are normally distributed with mean = 68 kg, S.D. = 3 kg:






(i) Students with mass **greater than 72 kg**

$$Z = \frac{72 - 68}{3} = \frac{4}{3} = 1.33$$

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

Using symmetry:

$$P(Z > 1.33) = 0.5 - P(0 < Z < 1.33)$$

From Z-tables: $P(0 < Z < 1.33) = 0.4082$.

$$P(Z > 1.33) = 0.5 - 0.4082 = 0.0918$$

Number of students:

$$300 \times 0.0918 \approx 28$$

**Answer:** $P(Z > 1.33) = 0.0918 \;\Rightarrow\; 28$ students

---

(ii) Students with mass **less than or equal to 64 kg**

$$Z = \frac{64 - 68}{3} = \frac{-4}{3} = -1.33$$

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

Using symmetry:

$$P(Z \leq -1.33) = 0.5 - P(0 < Z < 1.33)$$

From Z-tables: $P(0 < Z < 1.33) = 0.4082$.

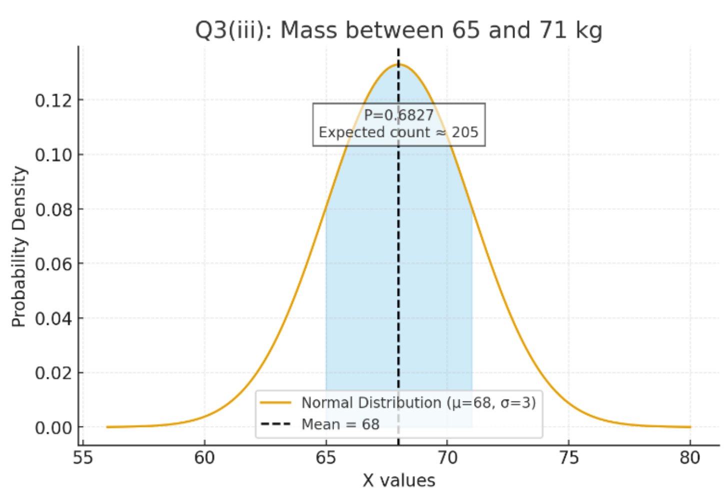
$$P(Z \leq -1.33) = 0.5 - 0.4082 = 0.0918$$

Number of students:

$$300 \times 0.0918 \approx 28$$

**Answer:** $P(Z < -1.33) = 0.0918 \;\Rightarrow\; 28$ students

---




(iii) Students with mass **between 65 kg and 71 kg inclusive**

For $X = 65$:

$$Z_1 = \frac{65 - 68}{3} = -1$$

For $X = 71$:

$$Z_2 = \frac{71 - 68}{3} = 1$$

$$P(65 \leq X \leq 71) = P(-1 \leq Z \leq 1)$$

Break into two parts:

$$P(-1 \leq Z \leq 1) = P(-1 \leq Z \leq 0) + P(0 \leq Z \leq 1)$$

From Z-tables:

- $P(0 < Z < 1) = 0.3413$

So,

$$P(-1 < Z < 0) = 0.3413, \quad P(0 < Z < 1) = 0.3413$$

$$P(-1 < Z < 1) = 0.3413 + 0.3413 = 0.6826$$

Number of students:

$$300 \times 0.6826 = 204.8 \approx 205$$

**Answer:** $P(-1 < Z < 1) = 0.6826 \;\Rightarrow\; 205$ students

---

**4.** (a) 1000 students have written an exam. The mean marks = 35 and S.D. = 5. Find:

- o How many students' marks lie between 25 and 40?
- o How many students got more than 40 marks?
- o How many students got below 20 marks?

(b) Find the mean and S.D. of a normal distribution in which 7% of items are under 35 and 89% are under 63.

---

**Solution: 4(a):** 1000 students have written an exam. Mean = 35, S.D. = 5. Find:

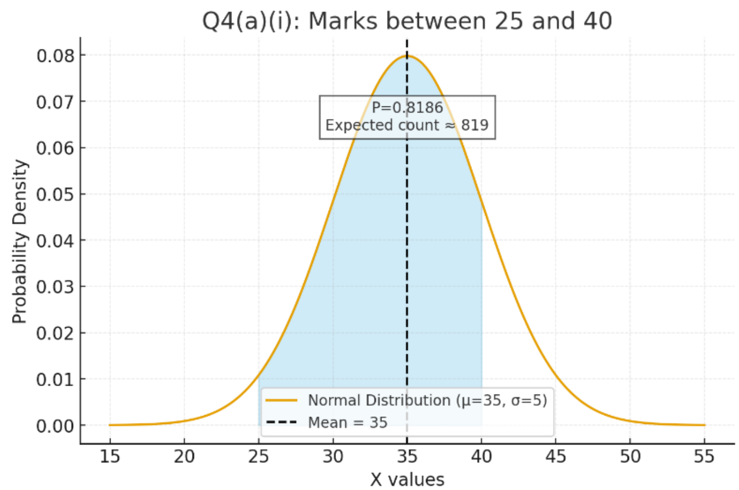
- How many students' marks lie between 25 and 40
- How many students got more than 40 marks
- How many students got below 20 marks

---

Write given values; $\mu = 35, \quad \sigma = 5, \quad N = 1000$

We will use the **Z-transformation**: $Z = \frac{X-\mu}{\sigma}$

---




(i) Students with marks between 25 and 40

For $X = 25$:

$$Z_1 = \frac{25 - 35}{5} = -2$$

For $X = 40$:

$$Z_2 = \frac{40 - 35}{5} = 1$$

So,

$$P(25 < X < 40) = P(-2 < Z < 1)$$

Break into two parts:

$$P(-2 < Z < 1) = P(-2 < Z < 0) + P(0 < Z < 1)$$

From Z-tables:

- $P(0 < Z < 2) = 0.4772 \;\Rightarrow\; P(-2 < Z < 0) = 0.4772$
- $P(0 < Z < 1) = 0.3413$

So,

$$P(-2 < Z < 1) = 0.4772 + 0.3413 = 0.8185$$

Number of students:

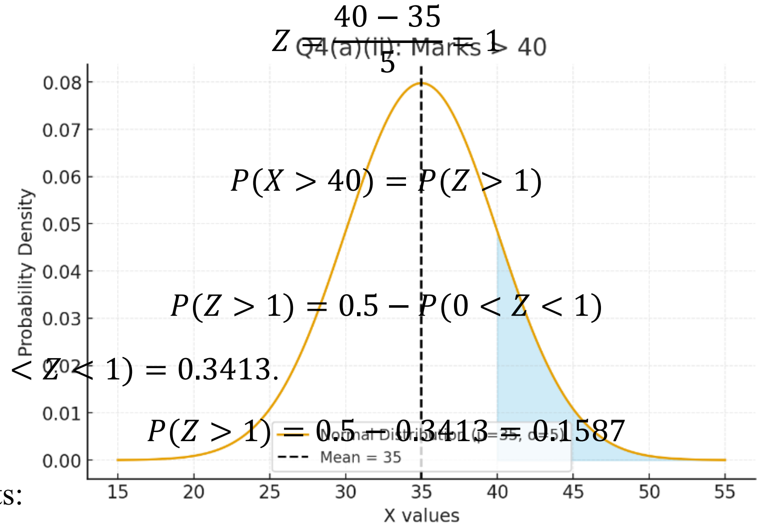
$$1000 \times 0.8185 = 819$$

**Answer:** $P(-2 < Z < 1) = 0.8185 \;\Rightarrow\; 819$ students

---

(ii) Students with marks more than 40

$$Z = \frac{40 - 35}{5} = 1$$




So,

$$P(X > 40) = P(Z > 1)$$

Using symmetry:

$$P(Z > 1) = 0.5 - P(0 < Z < 1)$$

From Z-table: $P(0 < Z < 1) = 0.3413$.

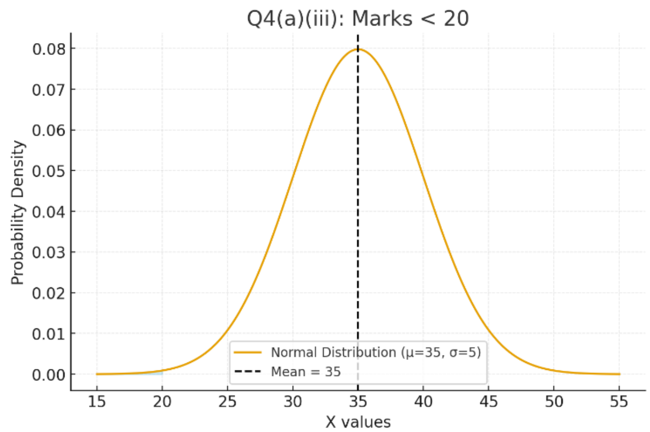
$$P(Z > 1) = 0.5 - 0.3413 = 0.1587$$

Number of students:

$$1000 \times 0.1587 = 159$$

**Answer:** $P(Z > 1) = 0.1587 \;\Rightarrow\; 159$ students

---




(iii) Students with marks below 20

$$Z = \frac{20 - 35}{5} = -3$$

So,

$$P(X < 20) = P(Z < -3)$$

Using symmetry:

$$P(Z < -3) = 0.5 - P(0 < Z < 3)$$

From Z-tables: $P(0 < Z < 3) = 0.49865$.

$$P(Z < -3) = 0.5 - 0.49865 = 0.00135$$

Number of students:

$$1000 \times 0.00135 \approx 1$$

**Answer:** $P(Z < -3) = 0.00135 \;\Rightarrow\; 1$ student

---

**Q4 (b):** Find the mean and S.D. of a normal distribution in which **7% of items are under 35** and **89% are under 63**.

---

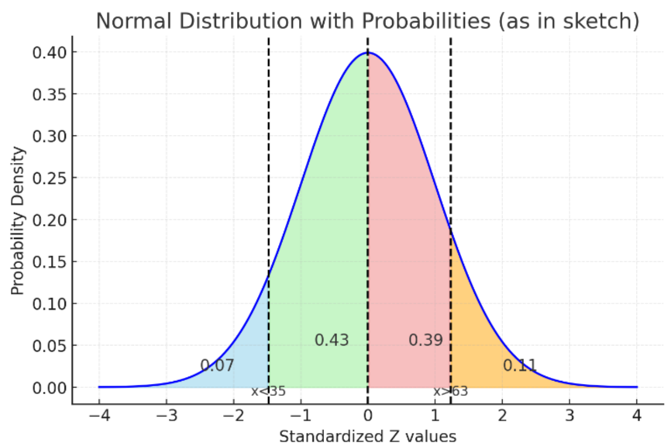
Find mean and S.D. of a normal distribution in which:

- 7% are below 35 → $P(X < 35) = 0.07$

- 89% are below 63 → $P(X < 63) = 0.89$

$\Rightarrow 1 - P(X < 63) = P(X > 63) = 1 - 0.89 = 0.11$

---




Step 1: Convert probabilities to Z-scores

- For probability = 0.07 ⇒ $0.5 - 0.07 = 0.43 \Rightarrow z_1 = -1.48$
- For probability = 0.11 ⇒ $0.5 - 0.11 = 0.39 \Rightarrow z_2 = 1.23$

For $P(Z < z) = 0.07$:

$$z_1 = -1.48 \quad (from\ normal\ tables)$$

For $P(Z < z) = 0.89$:

$$z_2 = 1.23 \quad (from\ normal\ tables)$$

So:

$$\frac{35 - \mu}{\sigma} = -1.48 \quad \dots (1)$$

$$\frac{63 - \mu}{\sigma} = 1.23 \quad \dots (2)$$

---

Step 2: Solve equations

From (1):

$$35 - \mu = -1.48\sigma \quad \Rightarrow \quad \mu = 35 + 1.48\sigma$$

From (2):

$$63 - \mu = 1.23\sigma \quad \Rightarrow \quad \mu = 63 - 1.23\sigma$$

Equating:

$$35 + 1.48\sigma = 63 - 1.23\sigma$$

$$2.71\sigma = 28 \quad \Rightarrow \quad \sigma = \frac{28}{2.71} \approx 10.33$$

Now substitute into (1):

$$\mu = 35 + 1.48(10.33) \approx 35 + 15.3 = 50.3$$

---

**Q5**(a) Find the mean and S.D. of a normal distribution in which 31% of items are under 45 and 8% are over 64.

(b) Suppose the weights of 800 male students are normally distributed with mean 28.8 kg and S.D. 2.06 kg. Find the number of students whose weights are

- between 28.4 kg and 30.4 kg
- more than 31.3 kg

---

**Solution(a):** Find the mean and S.D. of a normal distribution in which **31% of items are under 45** and **8% are over 64**.

---

Step 1: Write probability statements

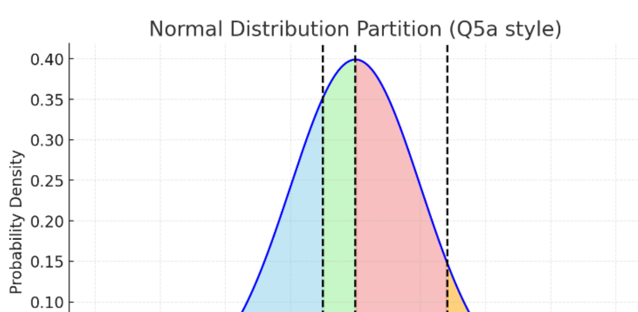
- $P(X < 45) = 0.31$
- $P(X > 64) = 0.08$

Let mean = $\mu$, standard deviation = $\sigma$.

---

Step 2: Convert to Z-scores using normal table

- For probability = 0.31 ⇒ $0.5 - 0.31 = 0.19 \Rightarrow \; z_1 = -0.50$
- For probability = 0.08 ⇒ $0.5 - 0.08 = 0.42 \Rightarrow z_2 = 1.41$



---

Step 3: Write equations

$$\frac{45-\mu}{\sigma} = -0.50 \quad \ldots(1)$$

$$\frac{64-\mu}{\sigma} = 1.41 \quad \ldots(2)$$

---

Step 4: Solve simultaneously

From (1):

$$45-\mu = -0.50\sigma \quad \Longrightarrow \quad \mu = 45 + 0.50\sigma$$

From (2):

$$64-\mu = 1.41\sigma \quad \Longrightarrow \quad \mu = 64 - 1.41\sigma$$

Equating:

$$45 + 0.50\sigma = 64 - 1.41\sigma$$

$$0.50\sigma + 1.41\sigma = 19$$

$$1.91\sigma = 19 \quad \Rightarrow \quad \sigma \approx 9.95$$

Now substitute back:

$$\mu = 45 + 0.50(9.95) \approx 45 + 4.98 = 49.98$$

---

**Solution(b):** Weights of 800 male students are normally distributed with:

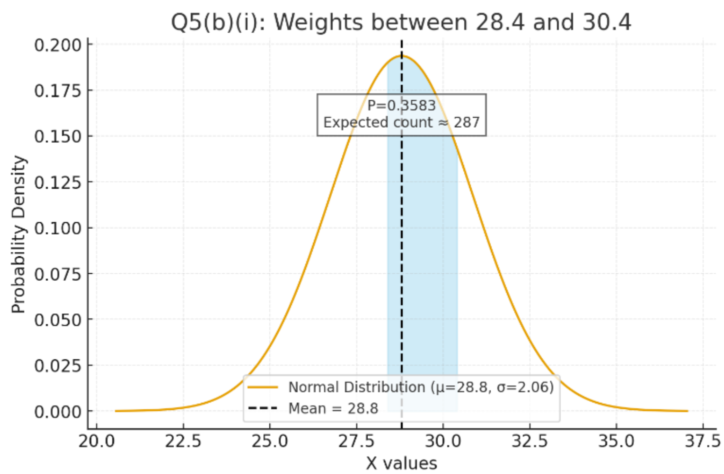
$$\mu = 28.8 \quad \sigma = 2.06 \quad N = 800$$

We need to find:

Number of students with weights between 28.4 and 30.4

Number of students with weights more than 31.3

---




**Case 1:** Between 28.4 and 30.4

For $X = 28.4$:

$$Z_1 = \frac{28.4 - 28.8}{2.06} = -0.19$$

For $X = 30.4$:

$$Z_2 = \frac{30.4 - 28.8}{2.06} = 0.78$$

So,

$$P(28.4 < X < 30.4) = P(-0.19 < Z < 0.78)$$

Break into two parts:

$$P(-0.19 < Z < 0.78) = P(-0.19 < Z < 0) + P(0 < Z < 0.78)$$

From Z-tables:

- $P(0 < Z < 0.19) = 0.0751 \;\Rightarrow\; P(-0.19 < Z < 0) = 0.0751$
- $P(0 < Z < 0.78) = 0.2822$

So,

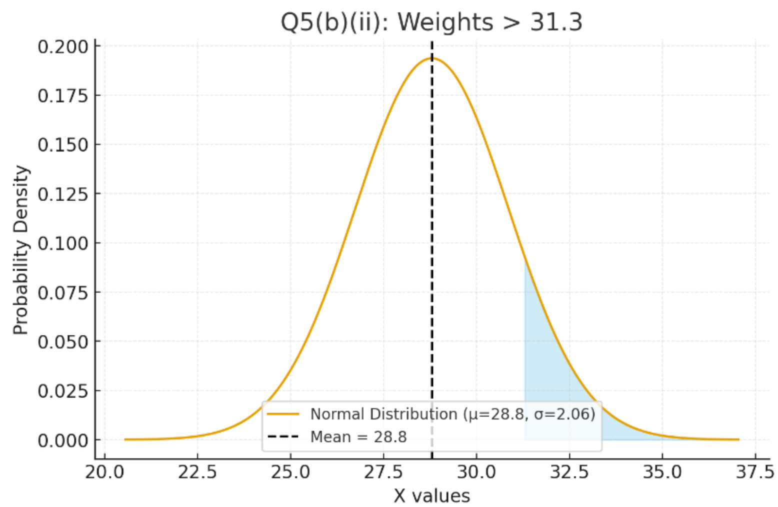
$$P(-0.19 < Z < 0.78) = 0.0751 + 0.2822 = 0.3573$$

Number of students:

$$800 \times 0.3573 \approx 286$$

**Answer:** $P(-0.19 < Z < 0.78) = 0.3573 \;\Rightarrow\; 286$ students

---




**Case 2:** More than 31.3

For $X = 31.3$:

$$Z = \frac{31.3 - 28.8}{2.06} = 1.21$$

So,

$$P(X > 31.3) = P(Z > 1.21)$$

Using symmetry:

$$P(Z > 1.21) = 0.5 - P(0 < Z < 1.21)$$

From Z-tables: $P(0 < Z < 1.21) = 0.3869$.

So,

$$P(Z > 1.21) = 0.5 - 0.3869 = 0.1131$$

Number of students:

$$800 \times 0.1131 \approx 91$$

**Answer:** $P(Z > 1.21) = 0.1131 \;\; \Rightarrow \;\; 91$ students

---

**Q6.** In a sample of 1000 cases the mean of a test = 14 and S.D. = 2.5. Assuming normal distribution, find:

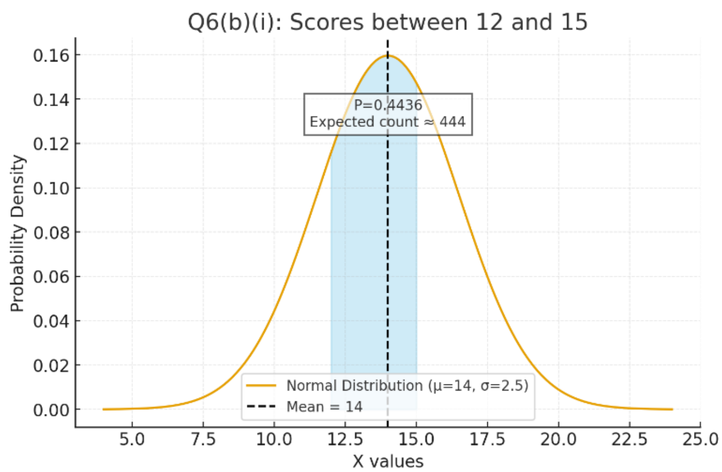
- How many students score between 12 and 15?
- How many students score above 18?

How many students score below 18?

---

**Solution:** Step 1: Write given data; $\mu = 14, \quad \sigma = 2.5, \quad N = 1000$

---




**(i) Students scoring between 12 and 15**

$$Z_1 = \frac{12 - 14}{2.5} = -0.8, \quad Z_2 = \frac{15 - 14}{2.5} = 0.4$$

So,

$$P(12 < X < 15) = P(-0.8 < Z < 0.4)$$

Break into two parts:

$$P(-0.8 < Z < 0.4) = P(-0.8 < Z < 0) + P(0 < Z < 0.4)$$

From Z-tables:

- $P(0 < Z < 0.8) = 0.2881 \Rightarrow P(-0.8 < Z < 0) = 0.2881$
- $P(0 < Z < 0.4) = 0.1554$

So,

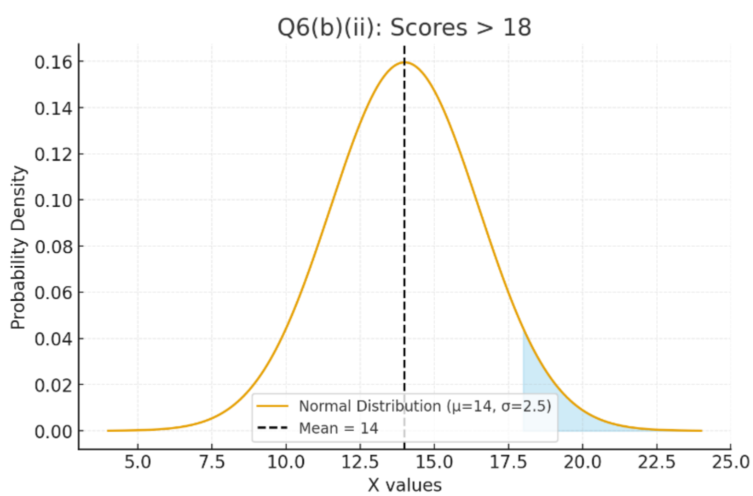
$$P(-0.8 < Z < 0.4) = 0.2881 + 0.1554 = 0.4435$$

Number of students:

$$1000 \times 0.4435 = 444$$

**Answer:** $P(-0.8 < Z < 0.4) = 0.4435 \Rightarrow 444$ students

---




**(ii) Students scoring above 18**

$$Z = \frac{18 - 14}{2.5} = 1.6$$

So,

$$P(X > 18) = P(Z > 1.6)$$

By symmetry:

$$P(Z > 1.6) = 0.5 - P(0 < Z < 1.6)$$

From Z-tables: $P(0 < Z < 1.6) = 0.4452$.

So,

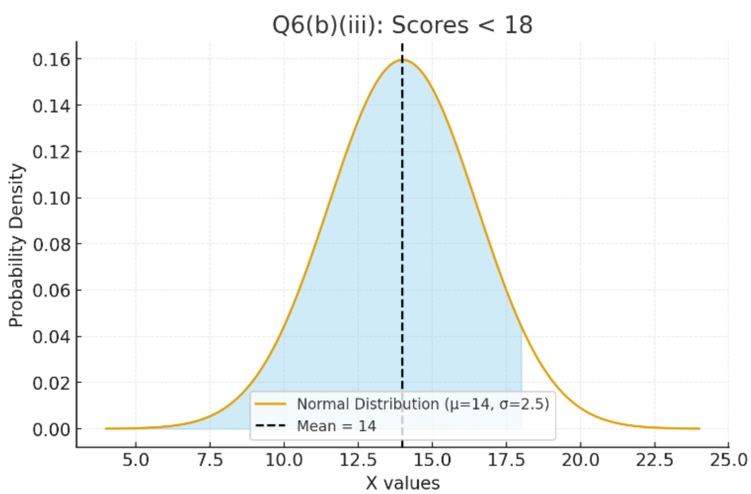
$$P(Z > 1.6) = 0.5 - 0.4452 = 0.0548$$

Number of students:

$$1000 \times 0.0548 = 55$$

**Answer:** $P(Z > 1.6) = 0.0548 \;\Rightarrow\; 55$ students

---




**(iii) Students scoring below 18**

$$P(X < 18) = P(Z < 1.6)$$

Now:

$$P(Z < 1.6) = 0.5 + P(0 < Z < 1.6)$$

$$= 0.5 + 0.4452 = 0.9452$$

Number of students:

$$1000 \times 0.9452 = 945$$

**Answer:** $P(Z < 1.6) = 0.9452 \;\Rightarrow\;$ 945 students

---

## Sampling distributions

**Introduction:**

Given a variable X, if we arrange its values in ascending order and assign probability to each of the values or if we present $X_i$ in a form of relative frequency distribution the result is called Sampling Distribution $of\ X$

**Definitions:**

**Population:** Collection of all objects or totality of the quantity or the group of individuals is called population. The other name for population is Universe.

A population consists of finite number of objects is called finite population otherwise it is infinite population.

**Size of the population:** The number of objects in a population is known as size of the population and it is denoted by N.

**Sample:** A finite subset of population is called a sample.

**Size of the sample:** The number of objects in a sample is known as size of the sample and it is denoted by n.

Here sample are two types.

1. Small sample 2. Large sample

1. Small sample: If the size of the sample is n < 30 is known as small sample
2. Large sample: If the size of the sample is n ≥ 30 is known as large sample

Sample size (n)

| | |
|---|---|
| Large samples (n ≥ 30) | Small sample (n < 30) |
| Population size(N) | Statistics (n) |
| (Parameter ($\mu$), Variance ($\sigma^2$), S.D ($\sigma$)) | (Mean ($\underline{x}$), Variance ($s^2$), S.D ($s$)) |

1. Parameter: Characteristic or measure obtained from a population.

2. Statistic: Characteristic or measure obtained from a sample.

3. Sampling: The process or method of sample selection from the population.

4. Sampling unit: the ultimate unit to be sampled or elements of the population to be sampled.

**Types of sampling distribution:** There are four types of sampling distributions

**1. Purposive sampling:** If the sample elements are selected with a definite purpose in view, then the sample selected is known as purposive sampling.

2. **Random sampling:** A sample in which every element of the population has an equal chance of being included

**3. Simple sampling** (Simple Random Sampling)**:** A method where each possible sample of a given size has an equal probability of selection, either with or without replacement.

**4. Stratified Random Sampling**: A method where the population is divided into homogeneous subgroups (strata), and random samples are drawn separately from each stratum.

**Formula:** Let N be the population size and n be the sample size then the number of samples can be drawn from the population with replacement is $N^n$ and without replacement is $N_{c_n}$

Ex: Let us consider a population of size 3 say {2,4,6} now draw a sample of size 2 with replacement and without replacement

Sol: given population = {2,4,6} = N = 3and sample size n = 2

Now, without replacement $N_{c_n} = 3_{c_2} = 3$

Consider sample of size 2 with replacement $N^n = 3^2 = 9$

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

**Note: Though the population is same the sample drawn from the same population may have different means.**

**Result: Mean of the sampling distribution of means$(\mu_{\underline{x}})$:**

Let $x_1, x_2, x_3, \ldots..x_n$ be the n samples which are drawn from a normal population with mean $\mu$ and variance$\sigma^2$. We know that its linear combination $\underline{x} = \frac{x_1+x_2+x_3,\ldots..+x_n}{n}$ is also a random variables and also we know that $E(x_1) = E(x_2) = E(x_3) = \cdots E(x_n) = \mu$

Now $\underline{x} = \frac{x_1+x_2+x_3,\ldots..+x_n}{n}$

$$\Rightarrow E(\underline{x}) = E\left(\frac{x_1+x_2+x_3,\ldots..+x_n}{n}\right)$$

$$= \frac{1}{n}E(x_1 + x_2 + x_3, \ldots..+x_n) = \frac{1}{n}E(x_1) + E(x_2) + E(x_3) + \ldots..+E(x_n)$$

$$= \frac{1}{n}(\mu + \mu + \cdots \mu(n - times)) = \frac{1}{n}(n\,\mu) = \mu$$

Therefore, $E(\underline{x}) = \mu_{\underline{x}} = \mu$

Hence the mean of sampling distribution of means is equal to population mean.

**Result: Variance of sampling distribution of mean**$(\sigma^2{}_{\underline{x}})$**:**

Let $x_1, x_2, x_3, \ldots..x_n$ be the n samples which are drawn from a normal population with mean $\mu$ and variance $\sigma^2$

we know that $V(x_1) = V(x_2) = V(x_3) = \cdots V(x_n) = \sigma^2$

Now, $V(\underline{x}) = V\left(\frac{x_1 + x_2 + x_3, \ldots.. + x_n}{n}\right) = \frac{1}{n^2} V(x_1 + x_2 + x_3, \ldots.. + x_n)$ [ Since V(ax) = $a^2 V(x)$]

$$= \frac{1}{n^2}\big(V(x_1) + V(x_2) + V(x_3), \ldots.. + V(x_n)\big) = \frac{1}{n^2}(\sigma^2 + \sigma^2 + \cdots + \sigma^2(n - times))$$

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

**Results:**

1. If the samples can be drawn from the population with replacement, then the following results are valid i.e $\mu_{\underline{x}} = \mu$ and $\sigma^2{}_{\underline{x}} = \frac{\sigma^2}{n}$
2. If the samples can be drawn from the population without replacement, then the following results are valid i.e $\mu_{\underline{x}} = \mu$ and $\sigma^2{}_{\underline{x}} = \left(\frac{N-n}{N-1}\right)\frac{\sigma^2}{n}$
   Here $\left(\frac{N-n}{N-1}\right)$ is called correction factor of the population
   where N = Population size and n = sample size
3. Standard error: S.D of the sampling distribution of means is known as Standard error
   It is denoted by $\sigma_{\underline{x}} = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$

**Problem1:** If the population is 3,6,9,15,27

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

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

c) Find the standard deviation of sampling distribution of means

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

Standard deviation of the population,

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

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

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

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

a) Sampling without replacement:

The total number of samples without replacement is $N_{C_n} = 5_{c_3} = 10$

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

b) Mean of the sampling distribution of means is

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

C) $\sigma^2{}_{\underline{x}} = \frac{(6-12)^2+(8-12)^2+(9-12)^2+(10-12)^2+(12-12)^2+(13-12)^2+(14-12)^2+(15-12)^2+(16-12)^2+(17-12)^2}{10}$

$$= \frac{120}{10} = 12$$

$$\sigma_{\underline{x}} = \sqrt{12} = 3.464$$

**Problem2:** *A* population consist of five numbers 2,3,6,8 and 11. Consider all possible samples of size two which can be drawn with replacement from this population. Find

a) The mean of the population

b) The standard deviation of the population

c) The mean of the sampling distribution of means and

d) The standard deviation of the sampling distribution of means

**Sol:**

a) Mean of the Population is given by

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

b) Variance of the population is given by

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

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

$$= \frac{16+9+0+4+25}{5} = 10.8$$

$$\therefore \sigma = 3.29$$

c) Sampling with replacement

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

List of all possible samples with replacement are

{(2,2),(2,3),(2,6),(2,8),(2,11),(3,2),(3,3)(3,6),(3,8),(3,11) (6,2)(6,3), (6,6)(6,8), (6,11)(8,2)(8,3)(8,6)(8,8)(8,1

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

{2,2.5,4,5,6.5,2.5,3,4.5,5.5,7,4,4.5,6,7,8.5,5.5,5,7,8,9.5,6.5,7,8.5,9.5,11}

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

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

d) The variance of the sampling distribution of means is obtained by subtracting the mean 6

from each number in sampling distribution of means and squaring the result, adding all 25

numbers thus obtained and dividing by 25.

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

$$\sigma_{\underline{x}} = \sqrt{54} = 2.32$$

**Problem3:** When a sample is taken from an infinite population, what happens to the standard error of the mean if the sample size is decreased from 800 to 200

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

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

Then S. $E_1 = \frac{\sigma}{\sqrt{800}} = \frac{\sigma}{20\sqrt{2}}$

When $n_1 is$ reduced to 200

let $n = n_2 = 200$

Then S. $E_2 = \frac{\sigma}{\sqrt{200}} = \frac{\sigma}{10\sqrt{2}}$

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

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

**Problem4:** The variance of a population is 2. The size of the sample collected from the population is 169. What is the standard error of mean

**Sol:** $n = The$ size ofthe sample $= 169$

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

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

**Problem5:** The mean height of students in a college is 155$cms$ and standard deviation is 15. What is the probability that the mean height of 36 students is less than 157 $cms$.

$Sol$: $\mu$ =Mean of the population = Mean height of students of a college $= 155cms$

$n$ = S.D ofpopulation $= 15cms$

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

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

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

$$= 0.5 + 0.2881 = 0.7881$$

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

**Problem6:** A random sample of size 100 is taken from a population with $\sigma = 5.1$. Given that the sample mean is $\underline{x} = 21.6$ Construct a 95% confidence limits for the population mean

**Sol:** Given $\underline{x} = 21.6$

$$Z\alpha/2 = 1.96, \qquad n = 100, \qquad \sigma = 5.1$$

Confidence interval = $(\underline{x} - Z\alpha/2(\frac{\sigma}{\sqrt{n}})\ ,\ \underline{x} + Z\alpha/2(\frac{\sigma}{\sqrt{n}}))$

$$\underline{x} - Z\alpha/2\left(\frac{\sigma}{\sqrt{n}}\right) = 21.6 - \frac{1.96 \times 5.1}{10} = 20.6$$

$$\underline{x} + Z\alpha/2\left(\frac{\sigma}{\sqrt{n}}\right) = 21.6 + \frac{1.96 \times 5.1}{10} = 22.6$$

Hence (20.6, 22.6) is the confidence interval for the population mean $\mu$

**Problem7:** Find the value of correction factor for n=50 and N = 300

**Sol:** Given n=50 and N = 300

We know that correction factor = $\frac{N-n}{N-1} = \frac{300-50}{300-1} = \frac{250}{299} = 0.836$

## Estimation Theory

Estimation theory helps us make decisions or draw conclusions about a population based on information from a sample.

<u>Estimate:</u> An **estimate** is a specific value used to approximate an unknown population parameter (like mean or variance).

<u>Estimator:</u> An **estimator** is the rule, method, or formula used to calculate that estimate

**<u>Types of estimation</u>**: There are two main types of estimation:

**1. Point Estimation:** When a single value is used to estimate a population parameter.
*Example:* Using the sample mean ($x^{-}$) to estimate the population mean (μ).

**2. Interval Estimation:** When a range of values (interval) is given, within which the parameter is likely to lie.

**Properties of estimation:** A good estimator should give results close to the true population value. The key properties are:

1. **Unbiasedness:**
   An estimator $\hat{\theta}$ is *unbiased* if its expected value equals the true parameter value.
   i.e., $E(\hat{\theta}) = \theta$
2. **Consistency:**
   An estimator is *consistent* if it gets closer to the true value $\theta$ as the sample size (n) increases.
   i.e., $\hat{\theta} \to \theta$ as $n \to \infty$

3. **Efficiency:**
   Between two unbiased estimators of the same parameter, the one with the smaller variance is *more efficient*.
   i.e., If $V(\hat{\theta}_1) < V(\hat{\theta}_2)$, then $\hat{\theta}_1$ is more efficientb) V $(\hat{\theta}_1)$ < V $(\hat{\theta}_2)$

**RESULT :1** Prove that for a random sample of size n ,$x_1$, $x_2$ …….$x_n$ taken from an infinite population $s^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \underline{x})^2$ is not unbiased estimator of the parameter $\sigma^2$ but $\frac{1}{n-1}\sum_{i=1}^{n} (x_i - \underline{x})^2$.

**Solution:** Let μ be the population mean.

Then $E(x_i) = \mu$ and $V(xi) = E[(x_i - \underline{x})]^2 = \sigma^2$ for I = 1,2,…….n

Now, $s^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \underline{x})^2$

$= \frac{1}{n}\sum_{i=1}^{n} (x_i{}^2 - 2x_i\underline{x} + \underline{x}^2) = \frac{1}{n}(\sum_{i=1}^{n} x_i{}^2 - 2\underline{x}\sum_{i=1}^{n} x_i + \underline{x}^2\sum_{i=1}^{n} 1)$

$= \frac{\sum x_i{}^2}{n} - 2\underline{x}^2 + \underline{x}^2$ (since $\sum_{i=1}^{n} x_i = n\,\underline{x}$ )

$= \frac{\sum x_i{}^2}{n} - \underline{x}^2$

$E(s^2) = \frac{1}{n}\sum E(x_i{}^2) - E(\underline{x}^2)$…….(1)

We know that $V(x) = E(x^2) - [E(x)]^2$ and $V(x_i) = E(x_i{}^2) - [E(x_i)]^2$

$\Rightarrow E(x_i{}^2) = V(x_i) + [E(x_i)]^2$

$\Rightarrow E(x_i{}^2) = \sigma^2 + \mu^2$……(2)

And also $E(\underline{x}^2) = V(\underline{x}) + [E(\underline{x})]^2$

$\Rightarrow E(\underline{x}^2) = \frac{\sigma^2}{n} + \mu^2$.....(3)

[Since $V(\underline{x})$ = **Variance of sampling distribution of mean**$(\sigma^2{}_{\underline{x}}) = \frac{\sigma^2}{n}$]

Substitute (2) and (3) in (1), we get

$E(s^2) = \frac{1}{n}[\sum (\sigma^2 + \mu^2)] - (\frac{\sigma^2}{n} + \mu^2)$

$E(s^2) = \frac{1}{n}(\sigma^2 + \mu^2)\sum 1 - (\frac{\sigma^2}{n} + \mu^2)$

$= \sigma^2 + \mu^2 - \frac{\sigma^2}{n} - \mu^2$ (since $\sum 1 = n$ for i=1to n)

$= \sigma^2(1 - \frac{1}{n})$

$\therefore$ E($s^2$) = $\sigma^2(\frac{n-1}{n})$......(4)

$\therefore$ E($s^2$) $\neq$ $\sigma^2$

Hence, $s^2 = \frac{1}{n}\sum_{i=1}^{n} \left(x_i - \underline{x}\right)^2$ is not unbiased estimator of the parameter $\sigma^2$

Let us consider, $S^2 = \frac{1}{n-1}\sum_{i=1}^{n} \left(x_i - \underline{x}\right)^2$

Multiply and divided with n

$$S^2 = \frac{1}{n-1} \times \frac{n}{n} \sum_{i=1}^{n} \left(x_i - \underline{x}\right)^2$$

$$= (\frac{n}{n-1})\frac{1}{n}\sum_{i=1}^{n} \left(x_i - \underline{x}\right)^2$$

$$= \left(\frac{n}{n-1}\right) \text{E } (s^2)$$

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

$\therefore$ E($S^2$) = $\sigma^2$

Hence $S^2 = \frac{1}{n-1}\sum_{i=1}^{n} \left(x_i - \underline{x}\right)^2$ is an unbiased estimator of the parameter $\sigma^2$

<u>**Interval estimation:**</u> In an interval estimation of the population parameter θ , if we can find two quantities $t_1$ and $t_2$ based on sample observations drawn from the population such that the unknown parameter θ is included in the interval ($t_1$ , $t_2$) in a specified percentage of case, then this interval is called a confidence interval for the parameter θ. The confidence interval has a specified confidence of correctly estimating the true value of the population parameter.

## Note

1. Standard error of mean ($\underline{x}$) $= \frac{\sigma}{\sqrt{n}}$

2. Standard error of Population(P) = $\sqrt{\frac{PQ}{n}}$ where P is population proportion and Q = 1 – P

3. Standard error of (Sample S.D) = $\frac{\sigma}{\sqrt{2\,n}}$

4. Standard error of (difference of two sample means) S.E $(\underline{x}_1 - \underline{x}_2) = \sqrt{\frac{{\sigma_1}^2}{n_1} + \frac{{\sigma_2}^2}{n_2}}$

5. Standard error of (difference of two proportions) S.E $(\underline{p}_1 - \underline{p}_2) = \sqrt{\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}}$

6. Standard error of (difference of two S.D) S.E $(s_1 - s_2) = \sqrt{\frac{{\sigma_1}^2}{2n_1} + \frac{{\sigma_2}^2}{2n_2}}$

**In case of large samples** $(n \geq 30)$

1. Maximum error of an estimate population mean $\mu = z_{\frac{\alpha}{2}}\left(\frac{\sigma}{\sqrt{n}}\right)$

2. M.E of population proportion = $z_{\frac{\alpha}{2}}\sqrt{\frac{PQ}{n}}$ where Q = 1 – P

**In case of small samples** $(n < 30)$

1. Maximum error of population mean $\mu = t_{\frac{\alpha}{2}}\left(\frac{s}{\sqrt{n}}\right)$ where $s = \sqrt{\frac{\Sigma \quad (x_i - \underline{x})^2}{n-1}}$

**Formula for confidence limits:**

**In case of large samples** $(n \geq 30)$**:**

1. The confidence interval for population mean $\mu = [\underline{x} - z_{\frac{\alpha}{2}}\left(\frac{\sigma}{\sqrt{n}}\right), \underline{x} + z_{\frac{\alpha}{2}}\left(\frac{\sigma}{\sqrt{n}}\right)]$

2. The confidence limits for difference of two means

= $[\left(\underline{x} - \underline{y}\right) - \{z_{\frac{\alpha}{2}}\sqrt{\frac{{\sigma_1}^2}{n_1} + \frac{{\sigma_2}^2}{n_2}}\}, \left(\underline{x} - \underline{y}\right) + \{z_{\frac{\alpha}{2}}\sqrt{\frac{{\sigma_1}^2}{n_1} + \frac{{\sigma_2}^2}{n_2}}\}]$

3. The confidence limits for population proportion P =$[P - z_{\frac{\alpha}{2}}\sqrt{\frac{PQ}{n}}, P + z_{\frac{\alpha}{2}}\sqrt{\frac{PQ}{n}}]$

4. The confidence limits for difference of two proportions

= $[(P_1 - P_2) - z_{\frac{\alpha}{2}}\sqrt{\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}} \ , (P_1 - P_2) + z_{\frac{\alpha}{2}}\sqrt{\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}}]$

If $P_1 \& P_2$ are unknown then we estimate population proportion as $P_1 = P_2 = P = \frac{n_1p_1 + n_2p_2}{n_1 + n_2}$

Ans Q = 1 – P then the confidence limits become

$[(P_1 - P_2) - z_{\frac{\alpha}{2}}\sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \ , (P_1 - P_2) + z_{\frac{\alpha}{2}}\sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}]$

**In case of small samples** $(n < 30)$

5. The confidence limits for population mean $= [\underline{x} - t_{\frac{\alpha}{2}}\left(\frac{s}{\sqrt{n}}\right), \underline{x} + t_{\frac{\alpha}{2}}\left(\frac{s}{\sqrt{n}}\right)]$ where $s = \sqrt{\frac{\Sigma \quad (x_i - \underline{x})^2}{n-1}}$

$t_{\frac{\alpha}{2}}$ is the t – value with $(n-1)$degree of freedom

6. The confidence limits for difference of two means

$$[(\underline{x}-\underline{y})-\{z_{\frac{\alpha}{2}}\sqrt{s^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}\},\ (\underline{x}-\underline{y})+\{z_{\frac{\alpha}{2}}\sqrt{s^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}\}]$$

Where $s=\sqrt{\frac{\Sigma\quad(x_i-\underline{x})^2+\Sigma\quad(y_i-\underline{y})^2}{n_1+n_2-2}}$ or $\frac{n_1{s_1}^2+n_2{s_2}^2}{n_1+n_2-2}$ $s_1, s_2$ are sample S.D

**Confidence limits for population mean μ in case of large samples:**

1) 95% Confidence limits are $\underline{x}\ \pm 1.96\,(S.E\ of\ \underline{x})$
2) 99% Confidence limits are $\underline{x}\ \pm 2.58\,(S.E\ of\ \underline{x})$
3) 90% Confidence limits are $\underline{x}\ \pm 1.64\ (S.E\ of\ \underline{x})$

**Confidence limits for the difference $\mu_1$- $\mu_2$ of two population means in case of large samples:**

1) 95% Confidence limits are $(\underline{x_1}-\underline{x_2})\ \pm 1.96\,(S.E\ of\ \underline{x_1}-\underline{x_2})$
2) 99% Confidence limits are $(\underline{x_1}-\underline{x_2})\ \pm 2.58(S.E\ of\ \underline{x_1}-\underline{x_2})$
3) 90% Confidence limits are $(\underline{x_1}-\underline{x_2})\ \pm 1.64\,(S.E\ of\ \underline{x_1}-\underline{x_2})$

**Confidence limits for population proportion P in case of large samples:**

1) 95% Confidence limits are $p\ \pm 1.96\,(S.E\ of\ p)$
2) 99% Confidence limits are $p\ \pm 2.58\,(S.E\ of\ p)$
3) 90% Confidence limits are $p\ \pm 1.64\ (S.E\ of\ p)$

**Confidence limits for the difference $P_1$- $P_2$ of two population proportions in case of large samples:**

1) 95% Confidence limits are $(p_1-p_2)\ \pm 1.96\,(S.E\ of\ p_1-p_2)$
2) 99% Confidence limits are $(p_1-p_2)\ \pm 2.58(S.E\ of\ p_1-p_2)$
3) 90% Confidence limits are $(p_1-p_2)\ \pm 1.64\,(S.E\ of\ p_1-p_2)$

**Confidence limits for population mean μ in case of small samples:**

(1-α) 100 % Confidence limits are $\underline{x}\ \pm t_{\frac{\alpha}{2}}\,(\frac{S}{\sqrt{n}})$ where $t_{\frac{\alpha}{2}}$ is the t- value with

$\vartheta$ = n-1 degrees of freedom, leaving an area $\frac{\alpha}{2}$ to the right .

**Confidence limits for the difference $\mu_1$- $\mu_2$ of two population means in case of small samples:**

(1-α) 100 % Confidence limits are $(\underline{x_1}-\underline{x_2})\ \pm t_{\frac{\alpha}{2}}S(\sqrt{\frac{1}{n_1}+\frac{1}{n_2}})$(where $t_{\frac{\alpha}{2}}$ is the t- value with $\vartheta$ = n-1 degrees of freedom , leaving an area $\frac{\alpha}{2}$ to the right .

where $S^2=\frac{n_1{s_1}^2+n_2{s_2}^2}{n_1+n_2-2}$ or $S^2=\frac{\sum_{i=1}^{n_1}\quad(x_i-\underline{x_i})^2+\sum_{j=1}^{n_2}\quad(x_j-\underline{x_j})^2}{n_1+n_2-2}$

**<u>Maximum error of estimate E for large sample:</u>** Since the sample mean estimate very rarely equals to the mean of population μ , a point estimate is generally accompanied with a statement of error which gives difference between estimate and the quantity to be estimated , thus the error is $|\underline{x} - \mu|$.

1) Maximum Error of an estimate μ ( in case of large samples) E = $Z_{\frac{\alpha}{2}}\left(\frac{\sigma}{\sqrt{n}}\right)$ and

   Sample size n = $\left(\frac{Z_{\frac{\alpha}{2}}\sigma}{E}\right)^2$

2) Maximum Error of an estimate P ( in case of large samples) = $Z_{\frac{\alpha}{2}}\sqrt{\frac{PQ}{n}}$

3) Maximum Error of an estimate μ ( in case of small samples) E = $t_{\frac{\alpha}{2}}\left(\frac{S}{\sqrt{n}}\right)$ , where $t_{\frac{\alpha}{2}}$ is the t- value with v= n-1 degrees of freedom , leaving an area $\frac{\alpha}{2}$ to the right .

**Problem1:** The mean and standard deviation of a population are 11,795and 14054 respectively. What can one assert with 95% confidence about the maximum error if $\underline{x}$ = 11795 and n = 50. And also construct 95% confidence interval for the true mean.

**Sol**: Here μ = 11795 ; σ = 14054 ; $\underline{x}$ = 11795 and n = 50

$Z_{\frac{\alpha}{2}}$ for 95% confidence = 1.96

Maximum Error of an estimate μ ( in case of large samples) E = $Z_{\frac{\alpha}{2}}\left(\frac{\sigma}{\sqrt{n}}\right)$

$$E = 1.96\left(\frac{14054}{\sqrt{50}}\right) = 3895.5743$$

95% Confidence limits are $\underline{x} \pm 1.96\,(S.E\ of\ \underline{x})$

$= (\underline{x} - 1.96\,(S.E\ of\ \underline{x})\ ;\ \underline{x} + 1.96\,(S.E\ of\ \underline{x})\ )$

$=(\underline{x} - 1.96\left(\frac{\sigma}{\sqrt{n}}\right)\ ;\ \underline{x} + 1.96\left(\frac{\sigma}{\sqrt{n}}\right)\ )$

=( 11795 – 3895.5743 ; 11795 + 3895.5743)

= ( 7899.4257 ; 15690.5743).

**Problem2:** Assume that σ = 20, hoe large a random sample be taken to assert with the probability 0.95 that the sample mean will not differ from the true mean by more than 3.0 points.

**Sol:** Given M.E = 3.0 and σ = 20

We have $Z_{\frac{\alpha}{2}}$ for 95% confidence = 1.96

We know that Sample size n = $\left(\frac{Z_{\frac{\alpha}{2}}\sigma}{E}\right)^2$ = $\left(\frac{1.96\ X\ 20}{3}\right)^2$ = 170.74 = 171 approximately

**Problem3:** In a study of an automobile insurance a random sample of 80 body repairs costs had a mean or Rs.472.36 and the S.D of Rs. 62.35. If $\underline{x}$ is used as appoint estimate to the true average repair costs, with what confidence we can assert that maximum error doesn't exceed Rs. 10.

**Sol:** Given that n= 80; $\underline{x}$= 472.36; σ = 62.35 and M.E = 10

We know that M.E(E) = $Z_{\frac{\alpha}{2}}\left(\frac{\sigma}{\sqrt{n}}\right)$

$$Z_{\frac{\alpha}{2}} = \frac{(E)\ \sqrt{n}}{\sigma} = \frac{(10)\ \sqrt{80}}{62.35} = 1.43$$

$Z_{\frac{\alpha}{2}}$ = 1.43 , The area when Z= 1.43 from tables is 0.4236.

$\frac{\alpha}{2}$ = 0.4236 ; α = 0.8472

Confidence = 84.72%

**Problem4:** Among 900 people in a state 90 are found to be chapati eaters. Construct 99% confidence interval for the true proportion.

**Sol:** Given x = 90 , n= 900

p= x/n = 90/900 = 0.1 and q= 1-p = 0.9

We have $Z_{\frac{\alpha}{2}}$ for 99% confidence = 2.58

99% Confidence limits are $p\ \pm 2.58\ (S.E\ of\ p)$

$= (p - 2.58\ (S.E\ of\ p)\ \ ; p + 2.58\ (S.E\ of\ p)\ \ )$

$= (p - 2.58\ (\sqrt{\frac{PQ}{n}})\ \ ; p + 2.58\ (\sqrt{\frac{PQ}{n}})\ \ )$

(0.07; 0.13).

**Problem5:** If we can assert with 95% that the maximum error IS 0.05 and P = 0.2 . Find the size of the sample.

**Sol:** Given P = 0.2 and Q = 0.8

E = 0.05 and $Z_{\frac{\alpha}{2}}$ for 95% confidence = 1.96

Maximum Error of an estimate P $= Z_{\frac{\alpha}{2}}\sqrt{\frac{PQ}{n}}$

$$0.05 = 1.96\sqrt{\frac{0.2\ X 0.8}{n}}$$

n= 246.

**Problem6:** A random sample of 300 shoppers at a super market includes 204 who regularly use cents off coupons. In another sample of 500 shoppers at super market includes 75 who regularly use cents off coupons. Construct 98% confidence interval for the probability that any one shopper at the super market, selected at random, will regularly use cents off coupons.

**Sol:** Here $n_1$ = 300, $n_2$ = 500, $x_1$ = 204, $x_2$ = 75

$p_1 = \frac{x_1}{n_1} = 0.68$; $p_2 = \frac{x_2}{n_2} = 0.15$

98% Confidence limits are $(p_1 - p_2) \pm 2.33\ (S.E\ of\ p_1 - p_2)$

= (0.62; 0.74)

**Problem7:** A sample of size 10 was taken from a population S. D of sample is 0.03. Find the maximum error with 99% confidence.

**Sol:** Given n= 10; S = 0.03

$t_{\frac{\alpha}{2}}$ at 99% with V= 9 d.o.f is 3.25

Maximum Error of an estimate μ (in case of small samples) E = $t_{\frac{\alpha}{2}}\left(\frac{S}{\sqrt{n}}\right)$ ,

$$E = \frac{3.25\ X\ 0.03}{\sqrt{10}} = 0.0308.$$

**Problem8:** Ten bearings made by a certain process have a mean diameter of 0.5060cm with a S.D of 0.0040 cm. Assume that the data may be Taken as a random sample from a normal distribution, construct 95% confidence interval for the actual average diameter of the bearings.

**Sol:** Given n= 10; S = 0.0040

$t_{\frac{\alpha}{2}}$ at 95% with V= 9 d.o.f is 2.262

95% Confidence limits are $\underline{x} \pm t_{\frac{\alpha}{2}}\left(\frac{S}{\sqrt{n}}\right)$ = (0.5031; 0.5089)

## Large samples

**Testing of Hypothesis:** In many cases we are

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

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

**PROCEDURE FOR TESTING A HYPOTHESIS:**

Test of Hypothesis involves the following steps:

**Statement of hypothesis:**

There are two types of hypothesis:

**Step1: Null hypothesis:** A definite statement about the population parameter. Usually a null hypothesis is written as **there is no difference**, denoted by $H_0$.

Ex. $H_0: \mu = \mu_0$

**Step2: Alternative hypothesis:** A statement which contradicts the null hypothesis is called alternative hypothesis. Usually an alternative hypothesis is written as **there is some difference,** denoted by $H_1$.

Setting of alterative hypotheses is very important to decide whether we should go for two-tailed test or one -tailed test, this should be depends upon the given question.

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

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

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

**Step3: Specification of level of significance:**

The L.O.S is denoted by $\alpha$ where $\alpha$ is the confidence with which we reject or accept the null hypothesis.

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

**Step 4: Identification of the test Statistic:**

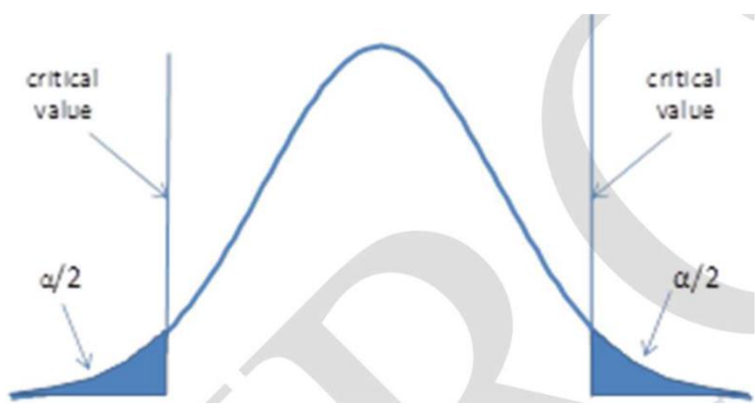
There are several tests of significance like $z, t, F$ etc. Depending upon the nature of the information given in the problem we have to select the right test and construct the test criterion and appropriate probability distribution

**Step5: Critical Region:**

It is the distribution of the statistic.

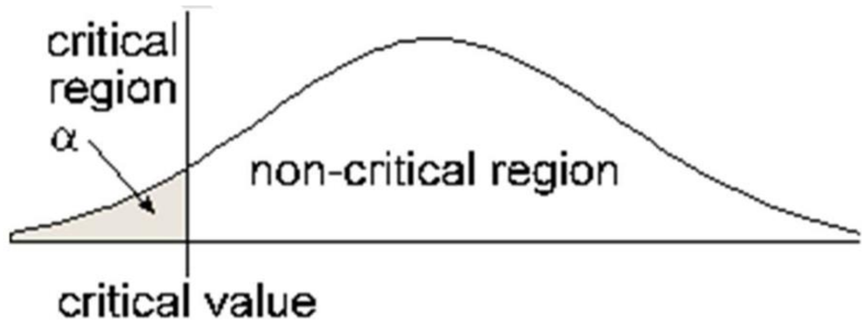
Two-Tailed Test: The critical region under the curve is equally distributed on both sides of the mean.

If $H_1$ has $\neq$sign, the critical region is divided equally on both sides of the distribution.

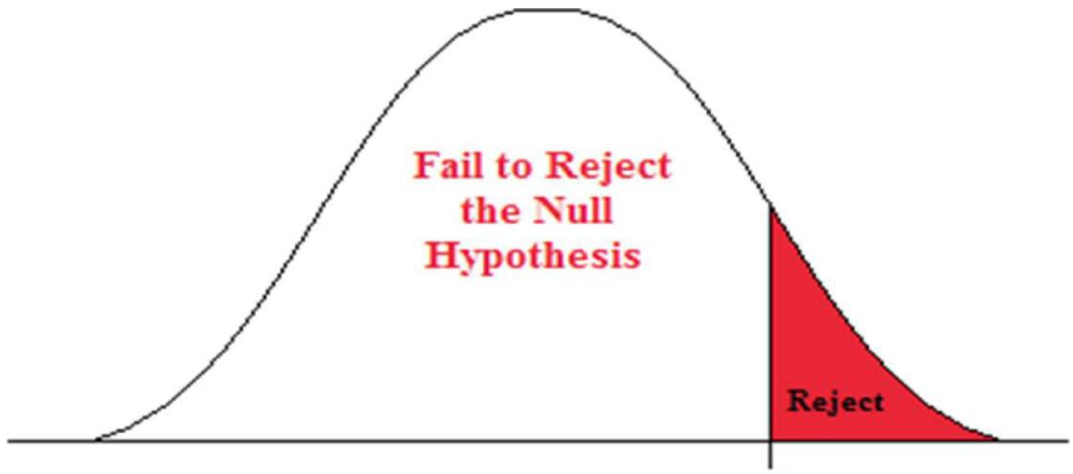



**One Tailed Test:** The critical region under the curve is distributed on one side of the mean.

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




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




Critical Value (+)

**Step6: Conclusion:**

By comparing the computed(calculated) value of $z_\alpha$ and the critical value (tabulated value) of $z_\alpha$ decision is taken for accepting or rejecting $H_0$

If calculated value of $z_\alpha \leq$ tabulated value of $z_\alpha$, we accept $H_0$, otherwise reject $H_0$.

**Errors of Sampling:**

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

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

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

**Large Samples:** Let a random sample of size $n \geq 30$ is defined as large sample

Applications of Large Samples

## 1.Test of Significance of a Single Mean

Let a random sample of size $n$, $\underline{x}$ $be$ the mean of the sample and $\mu$ be the population mean.

**Working Rule:**

1. **Null hypothesis:** $H_0$**:** There is no significant difference in the given population mean value say $\mu_0$ $i.e$ $H_0{:}\,\mu = \mu_0$

2. **Alternative hypothesis:** $H_1$**:** There is some significant difference in the given population mean value. i.e

$a)$ $H_1$ : $\mu \neq \mu_0$ (Two -tailed)

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

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

3. **Level of significance:** Set the LOS $\alpha$

4. **Test Statistic:** $z_{cal} = \frac{x-\mu}{\sigma/\sqrt{n}}$ (OR) $z_{cal} = \frac{x-\mu}{s/\sqrt{n}}$

5**: Critical Region:** Find the critical value (tabulated Value) $z_\alpha$ of z at the given level of significance

5. **Decision/conclusion**: If $z_{ca1}$value < $z_{tab}$ value accept $H_0$ otherwise reject $H_0$

**CRITICAL VALUES OF Z**

| LOS ∝ | 1% | 5% | 10% |
|---|---|---|---|
| $\mu \neq \mu_0$ | $\lvert Z\rvert$ >2.58 | $\lvert Z\rvert$>1.96 | $\lvert Z\rvert$>1.645 |
| $\mu > \mu_0$ | Z>2.33 | z>1.645 | Z>1.28 |
| $\mu < \mu_0$ | Z<-2.33 | Z<-1.645 | Z<-1.28 |

**NOTE:** Confidence limits for the mean of the population corresponding to the given sample.

$$\mu = \underline{x} \pm z_{\frac{\alpha}{2}}(S.E\ of\underline{X})\ i.e,$$

$\mu = \underline{x} \pm z_{\frac{\alpha}{2}}(\frac{\sigma}{\sqrt{n}})$ (or) $\mu = \underline{x} \pm z_{\frac{\alpha}{2}}(\frac{s}{\sqrt{n}})$

**Problems on Single Mean:**

**Problem1:** A sample of 64 students have a mean weight of 70 $kgs$. Can this be regarded as a sample mean from a population with mean weight 56 kgs and standard deviation 25$kgs$

**Sol:** Given $\underline{x}$ = mean of the sample = 70 kgs, $\mu$ =Mean of the population = 56 kgs,

$\sigma = S.D$ of population = 25 kgs, and $n$ =Sample size = 64

1. **Null Hypothesis** $H_0$**:** A Sample of 64 students with mean weight 70 kgs be regarded as a sample from a population with mean weight 56 kgs and standard deviation 25 $kgs$.

i.e., $H_0: \mu = 70$ kgs

2. **Alterative Hypothesis** $H_1$**:** Sample cannot be regarded as one coming from the population. i.e., $H_1: \mu \neq 70$ kgs (Two -tailed test)

3. **Level of significance** : $\alpha = 0.05 (Z_\alpha = 1.96)$

4. **Test Statistic:** $Z_{cal} = \frac{x-\mu}{\frac{\sigma}{\sqrt{n}}} = \frac{70-56}{\frac{25}{\sqrt{64}}} = 4.48$

5. **Critical Value:** 95% means in two – tailed test $\alpha = \frac{0.05}{2} = 0.025 \Rightarrow 0.5 - 0.25 = 0.475$ by table its value is 1.96 $\Rightarrow Z_{tab} = 1.96$

6.**Conclusion:** Since $|Z_{cal}|$ value > $Z_{tab}$value, we reject $H_0$

**Problem2:** In a random sample of 60 workers, the average time taken by them to get the work is 33.8 minutes with a standard deviation of 6.1 minutes. Can we reject the null hypothesis $\mu = 32.6$ in favor of alternative null hypothesis $\mu > 32.6$ at $\alpha = 0.05$(LOS)

**Sol:** Given $n = 60, \underline{x} = 33.8$, $\mu = 32.6$ and $\sigma = 6.1$

1. **Null Hypothesis** $H_0: \mu = 32.6$

2. **Alterative Hypothesis** $H_1: \mu > 32.6$ (Right one tailed test)

3. **Level of significance** : $\alpha = 0.05 (Z_\alpha = 1.64)$

4. **Test Statistic** : $Z_{cal} = \frac{x-\mu}{\frac{\sigma}{\sqrt{n}}} = \frac{33.8-32.6}{\frac{61}{\sqrt{60}}} = \frac{1.2}{0.7875} = 1.5238$

5. **Critical Value:** 95% means in one – tailed test $\alpha = 0.05 \Rightarrow 0.5 - 0.05 = 0.450$ by table its value is 1.64 $\Rightarrow Z_{tab} = 1.64$

6. **Conclusion**: Since $Z_{cal}$ value < $Z_{tab}$ value, we accept $H_0$

**Problem3:** A sample of 400 items is taken from a population whose standard deviation is 10. The mean of the sample is 40. Test whether the sample has come from a population with mean 38. Also calculate 95% confidence limits for the population.

**Sol:** Given $n = 400, \underline{x} = 40$, $\mu = 38$ and $\sigma = 10$

1. **Null Hypothesis** $H_0: \mu = 38$

2. **Alterative Hypothesis** $H_1: \mu \neq 38$ (Two -tailed test)

3. **Level of significance**: $\alpha = 0.05(Z_\alpha = 1.96)$

4. **Test statistic:** $Z_{cal} = \frac{\underline{x}-\mu}{\frac{\sigma}{\sqrt{n}}} = \frac{40-38}{\frac{10}{\sqrt{400}}} = \frac{2}{0.5} = 4$

5. **Critical Value:** 95% means in two – tailed test $\alpha = \frac{0.05}{2} = 0.025 \Rightarrow 0.5 - 0.25 = 0.475$ by table its value is 1.96 $\Rightarrow Z_{tab} = 1.96$

6. **Conclusion**: Since $|Z_{cal}| value > Z_{tab}$ value, we reject $H_0$

i.e., the sample is not from the population whose is 38.

95% confidence interval is $(\underline{x} - 1.96(\frac{\sigma}{\sqrt{n}}),\ \underline{x} + 1.96(\frac{\sigma}{\sqrt{n}}))$

i.e., $(40 - \frac{1.96(10)}{\sqrt{400}}\ ,\ 40 + \frac{1.96(10)}{\sqrt{400}}) = (40 - \frac{1.96(10)}{20},\ 40 + \frac{1.96(10)}{20})$

$= (40 - 0.98,\ 40 + 0.98) = (39.02,\ 40.98)$

**Problem4:** An ambulance service claims that it takes on the average less than 10 minutes to reach its destination in emergency calls. A sample of 36 calls has a mean of 11 minutes and the variance of 16 minutes. Test the claim at 0.05 l.o.s

**Sol:** Given $n = 36, \underline{x} = 11$, $\mu = 10$ and $\sigma = \sqrt{16} = 4$

1. **Null Hypothesis** $H_0 : \mu = 10$

2. **Alternate Hypothesis** $H_1: \mu < 10$ (Left one -tailed test)

3. **Level of significance**: $\alpha = 0.05(Z_\alpha = 1.645)$

4. **Test Statistic**: $Z_{cal} = \frac{\underline{x}-\mu}{\frac{\sigma}{\sqrt{n}}} = \frac{11-10}{\frac{4}{\sqrt{36}}} - = \frac{6}{4} = 1.5$

5. **Critical Value:** 95% means in one – tailed test $\alpha = 0.05 \Rightarrow 0.5 - 0.05 = 0.450$ by table its value is 1.64 $\Rightarrow Z_{tab} =$1.64

6. **Conclusion:** Since $|Z_{cal}|$ value $< Z_{tab}$ value, we accept $H_0$

### 2. Test of Significance for Difference of Means two Large Samples:

Let $\underline{x_1}$ & $\underline{x_2}$ *be* the means of the samples of two ramdom sizes $n_1$ & $n_2$ drawn from two populations having means $\mu_1$&$\mu_2$ and SD's $\sigma_1$ & $\sigma_2$

**Working Rule:**

1. **Null hypothesis:** $H_0$**:** $: \mu_1 = \mu_2$

**2. Alternative hypothesis:** $H_1$**:**

a) $H_1 : \mu_1 \neq \mu_2$ (Two Tailed) $b) H_1: \mu_1 < \mu_2$ (Left one tailed) $c) H_1: \mu_1 > \mu_2$ (Right one tailed)

**3. Level of Significance:** Set the LOS $\alpha$

**4. Test Statistic:** $Z_{cal} = \frac{(\underline{x}_1 - \underline{x}_2) - \delta}{SE\ of\ (\underline{x}_1 - \underline{x}_2)} = \frac{(\underline{x}_1 - \underline{x}_2) - \delta}{\sqrt{\frac{\sigma_1{}^2}{n_1} + \frac{\sigma_2{}^2}{n_2}}}$

Where $\delta = \mu_1 - \mu_2$ (where given constant). Other wise $\delta = \mu_1 - \mu_2 = 0$

if $\sigma_1^2 = \sigma_2^2 = \sigma^2$ then $Z_{cal} = \frac{(\underline{x}_1 - \underline{x}_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$

**5: Critical Region:** Find the critical value (tabulated Value) $z_\alpha$ of z at the given level of significance

**6. Conclusion:** if $|Z_{cal}| < Z_{tab}$, accept $H_0$ otherwise reject $H_0$

**NOTE:** Confidence limits for difference of means:

$$\mu_1 - \mu_2 = (\underline{x}_1 - \underline{x}_2) \pm z_{\alpha/2}\,[S.E\ of\ (\underline{x}_1 - \underline{x}_2)] = (\underline{x}_1 - \underline{x}_2) \pm z_{\alpha/2}\sqrt{\frac{\sigma_1{}^2}{n_1} + \frac{\sigma_2{}^2}{n_2}}$$

**Two – tailed test:**

**Problems on Difference of Means:**

**Problem1:** The means of two large samples of sizes 1000 and 2000 members are 67.5 inches and 68 inches respectively. Can the samples be regarded as drawn from the same population of S.D 2.5 inches.

**Sol:** Let $\mu_1$ and $\mu_2$ be the means of the two populations

Given $n_1 = 1000, n_2 = 2000$ and $\underline{x}_1 = 67.5$ inches, $\underline{x}_2 = 68$ inches Population $S.D$, $\sigma = 2.5$ inches

1. **Null Hypothesis** $H_0$: The samples have been drawn from the same population of S.D 2.5 inches i.e., $H_0: \mu_1 = \mu_2$

2. **Alternative Hypothesis** $: H_1: \mu_1 \neq \mu_2$ (Two-Tailed test)

3. **Level of significance**: $\alpha = 0.05$ ($z = 1.96$)

4. **Test Statistic**: $Z_{cal} = \frac{(\underline{x}_1 - \underline{x}_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{67.5 - 68}{(2.5)\sqrt{\frac{1}{1000} + \frac{1}{2000}}} = \frac{-0.5}{0.0968} = -5.16$

5**. Critical Region:** 95% means in two – tailed test $\alpha = \frac{0.05}{2} = 0.025 \Rightarrow 0.5 - 0.25 = 0.475$ by table its value is 1.96 $\Rightarrow Z_{tab} = 1.96$

6. **Conclusion**: Since $|Z_{cal}|$ value $> Z_{tab}$value, we reject $H_0$

Hence, we conclude that the samples are not drawn from the same population of S.D 2.5 inches.

**Problem2:** Samples of students were drawn from two universities and from their weights in kilograms, mean and standard deviations are calculated and shown below. Make a large sample test to test the significance of the difference between the means.

| | Mean | S. D | Size of the sample |
|---|---|---|---|
| University A | 55 | 10 | 400 |
| University B | 57 | 15 | 100 |

**Sol:** Let $\mu_1$ and $\mu_2$ be the means of the two populations

Given $n_1 = 400, n_2 = 100$ and $\underline{x}_1 = 55kgs,\ \underline{x}_2 = 57$ kgs, $\sigma_1 = 10\ and\ \sigma_2 = 15$

1. **Null Hypothesis** $H_0: \mu_1 = \mu_2$

2. **Alternative Hypothesis** $H_1: \mu_1 \neq \mu_2$ (Two-Tailed test)

3. **Level of significance**: $\alpha = 0.05(z = 1.96)$

4. **Test Statistic**: $Z_{cal} = \frac{(\underline{x}_1 - \underline{x}_2)}{\sqrt{\frac{\sigma_1{}^2}{n_1} + \frac{\sigma_2{}^2}{n_2}}} = \frac{55-57}{\sqrt{\frac{10^2}{400} + \frac{15^2}{100}}} = \frac{-2}{\sqrt{\frac{1}{4} + \frac{9}{4}}} = -1.26$

5**. Critical Region:** 95% means in two – tailed test $\alpha = \frac{0.05}{2} = 0.025 \Rightarrow 0.5 - 0.25 = 0.475$ by table its value is 1.96 $\Rightarrow Z_{tab}$ =1.96

6. **Conclusion**: Since $|Z_{cal}| value\ < Z_{tab}$ value, we accept $H_0$

Hence, we conclude that there is no significant difference between the means

**Problem3:** The average marks scored by 32 boys is 72 with a S.D of 8. While that for 36 girls is 70 with a S.D of 6. Does this data indicate that the boys perform better than girls at 5% los?

**Sol:** Let $\mu_1$ and $\mu_2$ be the means of the two populations

Given $n_1 = 32, n_2 = 36$ and $\underline{x}_1 = 72, \underline{x}_2 = 70,\ \sigma_1 = 8\ and\ \sigma_2 = 6$

1. **Null Hypothesis** $H_0: \mu_1 = \mu_2$

2. **Alterative Hypothesis** $H_1: \mu_1 > \mu_2$ (Right One Tailed test)

3. **Level of significance**: $\alpha = 0.05(Z_\alpha = 1.645)$

4. **Test Statistic**: $Z_{cal} = \frac{(\underline{x}_1 - \underline{x}_2)}{\sqrt{\frac{\sigma_1{}^2}{n_1}+\frac{\sigma_2{}^2}{n_2}}} = \frac{72-70}{\sqrt{\frac{8^2}{32}+\frac{6^2}{36}}} = \frac{2}{\sqrt{2+1}} = 1.1547$

5**. Critical Region:** 95% means in one – tailed test $\alpha = 0.05 \Rightarrow 0.5 - 0.05 = 0.450$ by table its value is 1.64 $\Rightarrow Z_{tab}$ =1.64

6. **Conclusion**: Since $|Z_{cal}|$ value $< Z_{tab}$ value, we accept $H_0$

Hence, we conclude that the performance of boys and girls is the same

**Problem4:** A sample of the height of 6400 Englishmen has a mean of 67.85 inches and a S.D of 2.56 inches while another sample of heights of 1600 Austrians has a mean of 68.55 inches and S.D of 2.52 inches. Do the data indicate that Austrians are on the average taller than the Englishmen? (Use $\alpha$ as 0.01)

**Sol:** Let $\mu_1$ and $\mu_2$ be the means of the two populations

Given $n_1 = 6400, n_2 = 1600$ and $\underline{x} = 67.85, \underline{x} = 68.55$, $\sigma_1 = 2.56 and \sigma_2 = 2.52$

1. **Null Hypothesis** $H_0 \colon \mu_1 = \mu_2$
2. **Alterative Hypothesis** $H_1 \colon \mu_1 < \mu_2$ (Left One Tailed test)
3. **Level of significance**: $\alpha = 0.01(Z_\alpha = -2.33)$

4.**Test Statistic**: $Z_{cal} = \frac{(\underline{x}_1 - \underline{x}_2)}{\sqrt{\frac{\sigma_1{}^2}{n_1}+\frac{\sigma_2{}^2}{n_2}}} = \frac{6785-6855}{\sqrt{\frac{(2.56)^2}{6400}+\frac{(2.52)^2}{1600}}}$

$$= \frac{67.85 - 68.55}{\sqrt{\frac{65536}{6400} + \frac{635}{1600}}}$$

$$= \frac{-0.7}{\sqrt{0.001 + 0.004}} = \frac{-0.7}{0.0707} = -9.9$$

5**. Critical Region:** 99% means in one – tailed test $\alpha = 0.01 \Rightarrow 0.5 - 0.01 = 0.490$ by table its value is 2.33 $\Rightarrow Z_{tab} = 2.33$

6. **Conclusion**: Since $|Z_{ca1}|$ value $> Z_{tab}$ value, we reject $H_0$

Hence, we conclude that Australians are taller than Englishmen.

**Problem5:** At a certain large university a sociologist speculates that male students spend considerably more money on junk food than female students. To test her hypothesis the sociologist randomly selects from records the names of 200 students. Of thee, 125 are men and 75 are women. The mean of the average amount spent on junk food per week by the men is Rs. 400 and S.D is 100. For the women the sample mean is Rs.450 and S.D is 150. Test the hypothesis at 5 % los?

**Sol:** Let $\mu_1$ and $\mu_2$ be the means of the two populations

Given $n_1 = 125, n_2 = 75$ and $\underline{x}_1$ = Mean of men = 400, $\underline{x}_2$ =Mean of women = 450

$$\sigma_1 = 100 \quad and \quad \sigma_2 = 150$$

1. **Null Hypothesis** $H_0: \mu_1 = \mu_2$

2. **Alternative Hypothesis** $H_1: \mu_1 > \mu_2$ (Right One Tailed test)

3. **Level of significance**: $\alpha = 0.05 (Z_\alpha = 1.645)$

4. **Test Statistic**: $Z_{cal} = \frac{(x_1 - \underline{x}_2)}{\sqrt{\frac{\sigma_1{}^2}{n_1} + \frac{\sigma_2{}^2}{n_2}}} = \frac{400-450}{\sqrt{\frac{100^2}{125} + \frac{150^2}{75}}} = \frac{-50}{\sqrt{80+300}}$

$$= \frac{-50}{\sqrt{380}} = \frac{-50}{19.49} = -2.5654$$

5**. Critical Region:** 95% means in one – tailed test $\alpha = 0.05 \Rightarrow 0.5 - 0.05 = 0.450$ by table its value is 1.64 $\Rightarrow Z_{tab} = 1.64$

6. **Conclusion**: Since $|Z_{ca1}|$ value $> Z_{tab}$ value, we reject $H_0$

**Problem6:** The research investigator is interested in studying whether there is a significant difference in the salaries of MBA grads in two cities. A random sample of size 100 from city A yields an average income of Rs. 20,250. Another random sample of size 60 from city $B$ yields an average income of Rs. 20,150. If the variance is given as $\sigma_1{}^2 = 40{,}000$ and $\sigma_2{}^2 = 32{,}400$ respectively. Test the equality of means and also construct 95% confidence limits.

**Sol:** Let $\mu_1$ and $\mu_2$ be the means of the two populations

Given $n_1 = 100, n_2 = 60$ and $\underline{x}_1$ = Mean of city $A$ = 20,250, $\underline{x}_2$ =Mean of city $B$ = 20,150

$\sigma_1{}^2 = 40{,}000$ and $\sigma_2{}^2 = 32{,}400$

1. **Null Hypothesis** $H_0: \mu_1 = \mu_2$

2. **Alternative Hypothesis** $H_1: \mu_1 \neq \mu_2$ (Two -Tailed test)

3. **Level of significance**: $\alpha = 0.05 (Z_\alpha = 1.96)$

4. **Test Statistic**: $Z_{cal} = \frac{(x_1 - \underline{x}_2)}{\sqrt{\frac{\sigma_1{}^2}{n_1} + \frac{\sigma_2{}^2}{n_2}}} = \frac{20{,}250 - 20{,}150}{\sqrt{\frac{40000}{100} + \frac{32400}{60}}} = \frac{100}{\sqrt{400+540}} = \frac{100}{30.66} = 3.26$

5**. Critical Region:** 95% means in two – tailed test $\alpha = \frac{0.05}{2} = 0.025 \Rightarrow 0.5 - 0.25 = 0.475$ by table its value is 1.96 $\Rightarrow Z_{tab} = 1.96$

6. **Conclusion**: Since $Z_{ca1}$value $> Z_{tab}$ value, we reject $H_0$

Hence, we conclude that there is a significant difference in the salaries of MBA grades two cities.

95% confidence interval is $\mu_1 - \mu_2 = (\underline{x_1} - \underline{x_2}) \pm 1.96\sqrt{\frac{\sigma_1{}^2}{n_1} + \frac{\sigma_2{}^2}{n_2}}$

$$= (20{,}150 - 20{,}250)) \pm 1.96\sqrt{\frac{40000}{100} + \frac{32400}{60}} = (39.90, 160.09)$$

### 3. Test of Significance for Single Proportions

Suppose a random sample of size $n$ has a sample proportion $p$ of members possessing a certain attribute (proportion of successes). To test the hypothesis that the proportion $P$ in the population has a specified value $P_0$

**Working Rule:**

1. **Null hypothesis:** $H_0$**:** $P = P_0$

2. **Alternative hypothesis**: $H_1$

a) $H_1 : P \neq P_0$ (Two Tailed test)

$b)H_1 : P < P_0$ (Left one-tailed)

c) $H_1 : P > P_0$ (Right one tailed)

3. **Level of significance:** At specified $L.O.S\ \alpha$, critical value of $Z$

4. **Test statistic:** $Z_{cal} = \frac{p-P}{\sqrt{\frac{PQ}{n}}}$ when $P$ is the Population proportion $Q = 1 - P$

**5: Critical Region:** Find the critical value (tabulated Value) $z_\alpha$ of z at the given level of significance

6. **Conclusion/ Decision**: if $|z_{cal}| < Z_{Tab}$, accept $H_0$ otherwise reject $H_0$

**NOTE:** Confidence limits for population proportion

$P = P \pm z_{\alpha/2}(\ S.E$ of P) $= P \pm z_{\alpha/2}(\sqrt{\frac{pq}{n}})$

**Problems on Single Proportion:**

**Problem1:** A die was thrown 9000 times and of these 3220 yielded a 3 or 4. Is this consistent with the hypothesis that the die was unbiased?

**Sol:** Given $n = 9000$

$P$ = Population of proportion of successes

$= P$ (getting a 3 or 4) $= \frac{1}{6} + \frac{1}{6} = \frac{2}{6} = 0.3333$

$Q = 1 - P = 0.6667$

$P$ = Proportion of successes of getting 3 or 4 in 9000 times $\frac{3220}{9000} = 0.3578$

1. **Null Hypothesis** $H_0$ : The die is unbiased

i.e., $H_0 : P = 0.33$

2. **Alternative Hypothesis** $H_1$ : The die is biased

i.e., $H_1$ : $P \neq 0.33$ (Two -Tailed test)

3. **Level of Significance** : $\alpha = 0.05 (Z_\alpha = 1.96)$

4. **Test Statistic**:$Z_{cal} = \frac{p-P}{\sqrt{\frac{PQ}{n}}} = \frac{0.3578-0.3333}{\sqrt{\frac{(0.3333)(0.6667)}{9000}}} = 4.94$

5**: Critical Region:** 95% means in two – tailed test $\alpha = \frac{0.05}{2} = 0.025 \Rightarrow 0.5 - 0.25 = 0.475$ by table its value is 1.96 $\Rightarrow Z_{tab} = 1.96$

6. **Conclusion**: Since $Z_{ca1}$value $> Z_{tab}$ value, we reject $H_0$

Hence, we conclude that the die is biased.

**Problem2:** In a random sample of 125 cool drinkers, 68 said they prefer thumbs up to Pepsi. Test the null hypothesis $P = 0.5$ against the alternative hypothesis hypothesis $P > 0.5$?

**Sol:** Given $n = 125$, $x = 68$ and $p = \frac{\chi}{n} = \frac{68}{125} = 0.544$

1. **Null Hypothesis** $H_0$ : $P = 0.5$

2. **Alternative Hypothesis** $H_1$ : $P > 0.5$( Right One Tailed test)

3. **Level of Significance**: $\alpha = 0.05 (Z_\alpha = 1.645)$

4. **Test Statistic**:$Z_{cal} = \frac{p-P}{\sqrt{\frac{PQ}{n}}} = \frac{0.544-0.5}{\sqrt{\frac{(0.5)(0.5)}{125}}} = 0.9839$

5**: Critical Region:** 95% means in two – tailed test $\alpha = \frac{0.05}{2} = 0.025 \Rightarrow 0.5 - 0.25 = 0.475$ by table its value is 1.96 $\Rightarrow Z_{tab} = 1.96$

6. **Conclusion**: Since $Z_{ca1}$value $< Z_{tab}$ value, we accept $H_0$

**Problem3:** A manufacturer claimed that at least 95% of the equipment which he supplied to a factory conformed to specifications. An experiment of a sample of 200 piece of equipment revealed that 18 were faulty. Test the claim at 5% los?

**Sol:** Given $n = 200$

Number of pieces confirming to specifications $= 200 - 18 = 182$

$p =$ Proportion of pieces confirming to specification $= \frac{182}{200} = 0.91$

$P =$ Population proportion $= \frac{95}{100} = 0.95$

1. **Null Hypothesis** $H_0 : P = 0.95$

2. **Alternative Hypothesis** $H_1 : P < 0.95$( Left One Tailed test)

3. **Level of Significance** $: \alpha = 0.05(Z_\alpha = -1.645)$

4. **Test Statistic**: $Z_{cal} = \frac{p-P}{\sqrt{\frac{PQ}{n}}} = \frac{0.91-0.95}{\sqrt{\frac{0.95\times 00591}{200}}} = -2.59$

5**: Critical Region:** 95% means in one – tailed test $\alpha = 0.05 \Rightarrow 0.5 - 0.05 = 0.450$ by table its value is $1.64 \Rightarrow Z_{tab} = 1.64$

6. **Conclusion:** Since $Z_{ca1}$value $> Z_{tab}$ value, we reject $H_0$.

Hence, we conclude that the manufacturer's claim is rejected.

**Problem4:** Among 900 people in a state 90 are found to be chapati eaters. Construct 99% confidence interval for the true proportion and also test the hypothesis for single proportion?

**Sol:** Given $x = 90$, $n = 900$

$$p = \frac{\chi}{n} = \frac{90}{900} = \frac{1}{10} = 0.1 \quad \text{and} \quad p = 0.99, \text{Q} = 1 - P = 1 - 0.99 = 0.01$$

Now $\sqrt{\frac{PQ}{n}} = \sqrt{\frac{(0.01)(0.99)}{900}} = \sqrt{\frac{0.0099}{900}} = \sqrt{0.000011} = 0.00331$

1. **Null Hypothesis** $H_0 : P = 0.5$

2. **Alternative Hypothesis** $H_1 : P \neq 0.5$( $Two$ Tailed test)

3. **Level of Significance** $: \alpha = 0.01(Z_\alpha = 2.58)$

4. **Test Statistic**:$Z_{cal} = \frac{p-P}{\sqrt{\frac{PQ}{n}}} = \frac{0.1-0.5}{\sqrt{\frac{0.99\times 0.01}{900}}} = \frac{-0.4}{0.00331} = -120.84$

5**: Critical Region:** 99% means in two – tailed test $\alpha = \frac{0.01}{2} = 0.005 \Rightarrow 0.5 - 0.005 = 0.495$ by table its value is 2.58 $\Rightarrow Z_{tab} = 2.58$

6. **Conclusion**: Since $|Z_{cal}| value > Z_{tab}$value, we reject $H_0$

Now, Confidence interval is $P = p \pm z_{\frac{\alpha}{2}}(\sqrt{\frac{pq}{n}})$

i.e., (0.1-0.00085398, 0.1+0.00085398) = $(0.0991, 0.1008)$

**4. Test for Significance for difference of Two Proportions (Populations)**

Let $p_1$ and $p_2$ be the sample proportions in two large random samples of sizes $n_1$& $n_2$ drawn from two populations having proportions $P_1$ & $P_2$

**Working Rule:**

1. **Null hypothesis:** $H_0: P_1 = P_2$

2. **Alternative hypothesis:**

a) $H_1: P_1 \neq P_2$ (Two Tailed)

$b) H_1: P_1 < P_2$ (Left one tailed)

c) $H_1: P_1 > P_2$ (Right one tailed)

**3. Level of significance:** At specified $LOS\ \alpha$ critical value of $Z$

4. **Test statistic** $Z_{cal} = \frac{(P_1 - P_2) - (P_1 - P_2)}{\sqrt{\frac{P_1 Q_1}{n_1} + \frac{P_1 Q_1}{n_2}}}$ if $(P_1 - P_2)$ is given.

If given only sample proportions then

$$Z_{cal} = \frac{p_1 - p_2}{\sqrt{\frac{P_1 q_1}{n_1} + \frac{P_2 q_1}{n_2}}} \quad \text{where } p_1 = \frac{x_1}{n_1} \ \& \ p_2 = \frac{x_2}{n_2}$$

OR

$$Z_{ca1} = \frac{p_1 - p_2}{\sqrt{pq(\frac{1}{n_1} + \frac{1}{n_2})}} \text{ Where } p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{x_1 + x_2}{n_1 + n_2} \text{ and } q = 1 - p$$

**5: Critical Region:** Find the critical value (tabulated Value) $z_\alpha$ of $z$ at the given level of significance

6. **Conclusion/Decision:** if $|Z_{cal}| < Z_{Tab}$, accept $H_0$ otherwise reject $H_0$

**NOTE:** Confidence limits for difference of population proportions

$$P_1 - P_2 = (p_1 - p_2) \pm z_{\alpha/2}\,(S.E\ of\ P_1 - P_2)$$

**Problems on Difference of Proportions:**

**Problem1:** Random samples of 400 men and 200 women in a locality were asked whether they would like to have a bus stop a bus stop near their residence. 200 men and 40 women in favor of the proposal. Test the significance between the difference of two proportions at 5% los?

**Sol:** Let $P_1$ and $P_2$ be the population proportions in a locality who favor the bus stop

Given $n_1$ =Number of men = 400

$n_2$ =number of women = 200

$x_1$ =Number of men in favor of the bus stop = 200

$x_2$ =Number of women in favor of the bus stop 40

$p_1 = \frac{x_1}{n_1} = \frac{200}{400} = \frac{1}{2}$ and $p_2 = \frac{x_2}{n_2} = \frac{40}{200} = \frac{1}{5}$

1. **Null Hypothesis** $H_0 : P_1 = P_2$

2. **Alternative Hypothesis** $H_1 : P_1 \neq P_2$ (Two Tailed test)

3. **Level of Significance** : $\alpha = 0.05 (Z_\alpha = 1.96)$

4. **Test Statistic**: $Z_{cal} = \frac{p_1 - p_2}{\sqrt{PQ(\frac{1}{n_1} + \frac{1}{n_2})}}$

We have $P = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{200+40}{400+200} = \frac{240}{600} = \frac{2}{5}$ and $Q = 1 - P = \frac{3}{5}$

$$= \frac{0.5 - 0.2}{\sqrt{(0.4)(0.6)(\frac{1}{400} + \frac{1}{200})}} = 7.07$$

5: **Critical Region:** 95% means in two – tailed test $\alpha = \frac{0.05}{2} = 0.025 \Rightarrow 0.5 - 0.25 = 0.475$ by table its value is $1.96 \Rightarrow Z_{tab} = 1.96$

6. **Conclusion**: Since $|Z_{ca1}|$ value > $Z_{tab}$ value, we reject $H_0$

Hence, we conclude that there is difference between the men and women in their attitude towards the bus stop near their residence.

**Problem2:** A machine puts out 16 imperfect articles in a sample of 500 articles. After the machine is overhauled it puts out 3 imperfect articles in a sample of100 articles. Has the machine been improved?

**Sol:** Let $P_1$ and $P_2$ be the proportions of imperfect articles in the proportion of articles manufactured by the machine before and after overhauling, respectively.

Given $n_1$ =Sample size before the machine overhauling = 500

$n_2$ =Sample size after the machine overhauling = 100

$x_1$ =Number of imperfect articles before overhauling = 16

$x_2$ =Number of imperfect articles after overhauling = 3

$$p_1 = \frac{x_1}{n_1} = \frac{16}{500} = 0.032 \; and \; p_2 = \frac{x_2}{n_2} = \frac{3}{100} = 0.03$$

1. **Null Hypothesis** $H_0 : P_1 = P_2$

2. **Alternative Hypothesis** $H_1 : P_1 > P_2$ (Right one Tailed test)

3. **Level of Significance** : $\alpha = 0.05 (Z_\alpha = 1.645)$

4. **Test Statistic**: $Z_{cal} = \frac{p_1 - p_2}{\sqrt{PQ(\frac{1}{n_1}+\frac{1}{n_2})}}$

We have $P = \frac{n_1p_1+n_2p_2}{n_1+n_2} = \frac{x_1+x_2}{n_1+n_2} = \frac{16+3}{500+100} = \frac{19}{600} = 0.032$

$$Q = 1 - p = 0.968$$

$$Z_{cal} = \frac{p_1 - p_2}{\sqrt{PQ(\frac{1}{n_1} + \frac{1}{n_2})}} = \frac{0.032 - 0.03}{\sqrt{(0.032)(0.968)(\frac{1}{500} + \frac{1}{100})}}$$

$$= \frac{0.002}{0.019} = 0.104$$

5**: Critical Region:** 95% means in one – tailed test $\alpha = 0.05 \Rightarrow 0.5 - 0.05 = 0.450$ by table its value is $1.64 \Rightarrow Z_{tab} = 1.64$

**6. Conclusion**: Since $|Z_{ca1}|$ value $< Z_{tab}$ value, we accept $H_0$

Hence, we conclude that the machine has improved.

**Problem3:** In an investigation on the machine performance the following results are obtained

| | No of units inspected | No of defectives |
|---|---|---|
| Machine 1 | 375 | 17 |
| Machine 2 | 450 | 22 |

Test whether there is any significant performance of two machines at $\alpha = 0.05$

**Sol:** Let $P_1$ and $P_2$ be the proportions of defective units in the population of units inspected in

machine 1 and Machine 2 respectively.

Given $n_1$ =Sample size ofthe Machine 1 = 375

$n_2$ = Sample size of the Machine 2 = 450

$x_1$ = Number of defectives of the Machine 1 = 17

$x_2$ = Number of defectives of the Machine 2 = 22

$$p_1 = \frac{x_1}{n_1} = \frac{17}{375} = 0.045 \; and \; p_2 = \frac{x_2}{n_2} = \frac{22}{450} = 0.049$$

1. **Null Hypothesis** $H_0 : P_1 = P_2$
2. **Alternative Hypothesis** $H_1 : P_1 \neq P_2$ (Two Tailed test)
3. **Level of Significance** : $\alpha = 0.05 (Z_\alpha = 1.96)$
4. **Test Statistic**: $Z_{cal} = \frac{p_1 - p_2}{\sqrt{PQ(\frac{1}{n_1}+\frac{1}{n_2})}}$

We have $P = \frac{n_1p_1+n_2p_2}{n_1+n_2} = \frac{x_1+x_2}{n_1+n_2} = \frac{17+22}{375+450} = \frac{39}{825} = 0.047$

$$Q = 1 - P = 1 - 0.047 = 0.953$$

$$Z_{cal} = \frac{p_1 - p_2}{\sqrt{PQ(\frac{1}{n_1}+\frac{1}{n_2})}} = \frac{0.045 - 0.049}{\sqrt{(0.047)(0.953)(\frac{1}{375}+\frac{1}{450})}} = -0.267$$

5**: Critical Region:** 95% means in two – tailed test $\alpha = \frac{0.05}{2} = 0.025 \Rightarrow 0.5 - 0.25 = 0.475$ by table its value is $1.96 \Rightarrow Z_{tab} = 1.96$

6. **Conclusion**: Since $|Z_{ca1}|$ value $< Z_{tab}$ value, we accept $H_0$

Hence, we conclude that there is no significant difference in performance of machines.

**Problem4:** A cigarette manufacturing firm claims that its brand A line of cigarettes outsells its brand $B$ by 8%. If it is found that 42 out of 200 smokers prefer brand A and 18 out of another sample of 100 smokers prefer brand B. Test whether 8% difference is a valid claim?

Sol: Given $n_1 = 200$ $n_2 = 100$

$x_1$ = Number of smokers preferring brand $A = 42$

$x_2$ = Number of smokers preferring brand $B = 18$

$$p_1 = \frac{x_1}{n_1} = \frac{42}{200} = 0.21 \ \ and \ \ p_2 = \frac{x_2}{n_2} = \frac{18}{100} = 0.18$$

And $\delta = P_1 - P_2 = 8\% = 0.08$

1. **Null Hypothesis** $H_0 : P_1 - P_2 = 0.08$

2. **Alternative Hypothesis** $H_1 : P_1 - P_2 \neq 0.08$ (Two – tailed test)

3. **Level of Significance** : $\alpha = 0.05 (Z_\alpha = 1.96)$

4. **Test Statistic**: $Z_{cal} = \frac{(p_1 - p_2) - \delta}{\sqrt{PQ(\frac{1}{n_1} + \frac{1}{n_2})}} = \frac{(p_1 - p_2) - (P_1 - P_2)}{\sqrt{PQ(\frac{1}{n_1} + \frac{1}{n_2})}}$

We have $P = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{42+1}{200+100} = \frac{60}{300} = 0.2$

$$Q = 1 - P = 1 - 0.2 = 0.8$$

$$Z_{cal} = \frac{(0.21 - 0.18) - 0.08}{\sqrt{(0.2)(0.8)(\frac{1}{200} + \frac{1}{100})}}$$

$$= \frac{-0.05}{0.0489} = -1.02$$

5**: Critical Region:** 95% means in two – tailed test $\alpha = \frac{0.05}{2} = 0.025 \Rightarrow 0.5 - 0.25 = 0.475$ by table its value is $1.96 \Rightarrow Z_{tab} = 1.96$

6. **Conclusion**: Since $|Z_{cal}|\ value < Z_{tab}$ value, we accept $H_0$

Hence, we conclude that 8% difference in the sale of two brands of cigarettes is a valid claim.

**Problem5:** In a city $A$, 20% of a random sample of 900 schoolboys has a certain slight physical defect. In another city B,18.5% of a random sample of 1600 school boys has the same defect. Is the difference between the proportions significant at 5% los?

**Sol:** Given $n_1 = 900$, $n_2 = 1600$

$x_1 = 20\%$ of $900 = 180$

$x_2 = 18.5\%$ of $1600 = 296,\ P_1 - P_2 = 18.5\% = 0.185$

$$p_1 = \frac{x_1}{n_1} = \frac{180}{900} = 0.2 \ \ and \ \ p_2 = \frac{x_2}{n_2} = \frac{296}{1600} = 0.185$$

1. **Null Hypothesis** $H_0 : P_1 = P_2$

2. **Alternative Hypothesis** $\underline{H_1}: P_1 \neq P_2$ (Two - tailed test)

3. **Level of Significance**: $\alpha = 0.05 (Z_\alpha = 1.96)$

4. **Test Statistic**: $Z_{cal} = \frac{(p_1 - p_2) - \delta}{\sqrt{PQ(\frac{1}{n_1} + \frac{1}{n_2})}} = \frac{(p_1 - p_2) - (P_1 - P_2)}{\sqrt{PQ(\frac{1}{n_1} + \frac{1}{n_2})}}$

We have $P = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{180+296}{900+1600} = \frac{476}{2500} = 0.19$

$$Q = 1 - P = 1 - 0.19 = 0.81$$

$$Z_{cal} = \frac{0.2 - 0.185}{\sqrt{(019)(081)(\frac{1}{900} + \frac{1}{1600})}}$$

$$= \frac{-0.015}{0.01634} = -0.918$$

5**: Critical Region:** 95% means in two – tailed test $\alpha = \frac{0.05}{2} = 0.025 \Rightarrow 0.5 - 0.25 = 0.475$ by table its value is $1.96 \Rightarrow Z_{tab} = 1.96$

6. **Conclusion**: Since $|Z_{cal}|\ value < Z_{tab}$ value, we accept $H_0$

Hence, we conclude that there is no significant difference between the proportions.

**Home work:**

**Problem1:** The mean life time of a sample of 100 light tubes produced by a company is found to be 1560 hrs with a population S.D of 90 $hrs$. Test the hypothesis for $\alpha = 0.05$ that the mean life time of the tubes produced by the company is 1580 $hrs$.

**Sol.** Given $\underline{x} = 1560$ hrs, $\mu = 1580 hrs,\ n = 100$ and $\sigma = 90 hrs$

1. **Null Hypothesis** $H_0 : H_0 : \mu = 1580$

2. **Alternative Hypothesis** $H_1 : \mu \neq 1580$ (Two -tailed test)

3. **Level of significance** : $\alpha = 0.05 (Z_\alpha = 1.96)$

4. **Test Statistic**: $Z_{cal} = \frac{x - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{1560 - 1580}{\frac{90}{\sqrt{100}}} = \frac{-20}{9}$

$$\therefore |z| = \frac{20}{9} = 2.22$$

5**: Critical Region:** 95% means in two – tailed test $\alpha = \frac{0.05}{2} = 0.025 \Rightarrow 0.5 - 0.25 = 0.475$ by table its value is 1.96 $\Rightarrow Z_{tab} = 1.96$

6. **Conclusion**: Since $|Z_{ca1}|$ value > $Z_{tab}$ value, we reject $H_0$

$$\therefore \mu \neq 1580$$

**Problem2:** Test the significance of the difference between the means of the samples from the following data

| | Sample A | Sample B |
|---|---|---|
| Size of sample | 100 | 150 |
| Mean | 50 | 51 |
| Standard Deviation | 4 | 5 |

(Table value = 1.96)

**Sol.** Let $\mu_1$ and $\mu_2$ be the means of the two populations

Given $n_1 = 100, n_2 = 150$ and $\underline{x} = 50, \underline{x} = 51$, $\sigma_1 = 4$ *and* $\sigma_2 = 5$

1. **Null Hypothesis** $H_0: \mu_1 = \mu_2$

2. **Alternative Hypothesis** $H_1: \mu_1 \neq \mu_2$ (Two-Tailed test)

3. **Level of significance**: $\alpha = 0.05 (Z_\alpha = 1.96)$

4. **Test Statistic** : $Z_{cal} = \frac{(\underline{x}_1 - \underline{x}_2)}{\sqrt{\frac{\sigma_1{}^2}{n_1} + \frac{\sigma_2{}^2}{n_2}}} = \frac{50-51}{\sqrt{\frac{4^2}{100} + \frac{5^2}{150}}} = \frac{-1}{\sqrt{\frac{16}{100} + \frac{25}{150}}} = \frac{-1}{\sqrt{0.16+0.167}} = -3.0581$

$$|Z_{cal}| = 3.0581$$

5**: Critical Region:** 95% means in two – tailed test $\alpha = \frac{0.05}{2} = 0.025 \Rightarrow 0.5 - 0.25 = 0.475$ by table its value is 1.96 $\Rightarrow Z_{tab} = 1.96$

6. **Conclusion**: Since $|Z_{cal}|$ value > $Z_{tab}$ value, we reject $H_0$

Hence, we conclude that there is some significant difference between the means

# CORRELATION AND REGRESSION

## CORRELATION

**Introduction:** In today's business world we come across many activities, which are dependent on each other. In businesses we see large number of problems involving the use of two or more variables. Identifying these variables and its dependency helps us in resolving the many problems. Many times, there are problems or situations where two variables seem to move in the same direction such as both are increasing or decreasing. At times an increase in one variable is accompanied by a decline in another. For example, family income and expenditure, price of a product and its demand, advertisement expenditure and sales volume etc. If two quantities vary in such a way that movements in one are accompanied by movements in the other, then these quantities are said to be correlated.

**Meaning:** Correlation is a statistical technique to ascertain the association or relationship between two or more variables. Correlation analysis is a statistical technique to study the degree and direction of relationship between two or more variables.

A correlation coefficient is a statistical measure of the degree to which changes to the value of one variable predict change to the value of another. When the fluctuation of one variable reliably predicts a similar fluctuation in another variable, there's often a tendency to think that means that the change in one causes the change in the other.

**Uses of correlations:**

1. Correlation analysis helps us in deriving precisely the degree and the direction of such relationship.

2. The effect of correlation is to reduce the range of uncertainty of our prediction. The prediction based on correlation analysis will be more reliable and near to reality.

3. Correlation analysis contributes to the understanding of economic behavior, aids in locating the critically important variables on which others depend, may reveal to the economist the connections by which disturbances spread and suggest to him the paths through which stabilizing forces may become effective

4. Economic theory and business studies show relationships between variables like price and quantity demanded advertising expenditure and sales promotion measures etc.

5. The measure of coefficient of correlation is a relative measure of change.

**Types of Correlation:**

Correlation is described or classified in several different ways. Three of the most important are:

I. **Positive and Negative** II. **Simple, Partial and Multiple** III. **Linear and non-linear**

---

1. **Positive and Negative Correlation**: Whether correlation is positive (direct) or negative (in-versa) would depend upon the direction of change of the variable.

**Positive Correlation**: If both the variables vary in the same direction, correlation is said to be positive. It means if one variable is increasing, the other on an average is also increasing or if one variable is decreasing, the other on an average is also deceasing, then the correlation is said to be positive correlation.

For example, the correlation between heights and weights of a group of persons is a positive correlation.

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Height(cm): X | 158 | 160 | 163 | 166 | 168 | 171 | 174 | 176 |
| Weight(kg): Y | 60 | 62 | 64 | 65 | 67 | 69 | 71 | 72 |

**Negative Correlation**: If both the variables vary in opposite direction, the correlation is said to be negative. If means if one variable increases, but the other variable decreases or if one variable decreases, but the other variable increases, then the correlation is said to be negative correlation. For example, the correlation between the price of a product and its demand is a negative correlation.

| | | | | | | |
|---|---|---|---|---|---|---|
| Price of product: X | 6 | 5 | 4 | 3 | 2 | 1 |
| Demand: Y | 75 | 120 | 175 | 215 | 250 | 400 |

---

**II: Simple, Partial and Multiple Correlation**:

The distinction between simple, partial and multiple correlation is based upon the number of variables studied.

**Simple Correlation**: When only two variables are studied, it is a case of simple correlation. For example, when one studies relationship between the marks secured by student and the attendance of student in class, it is a problem of simple correlation.

**Partial Correlation**: In case of partial correlation, one studies three or more variables but considers only two variables to be influencing each other and the effect of other influencing variables being held constant. For example, in above example of relationship between student marks and attendance, the other variable influencing such as effective teaching of teacher, use of teaching aid like computer, smart board etc are assumed to be constant.

**Multiple Correlation**: When three or more variables are studied, it is a case of multiple correlation. For example, in above example if study covers the relationship between student marks, attendance of students, effectiveness of teacher, use of teaching aids etc, it is a case of multiple correlation.

---

**III: Linear and Non-linear Correlation**:

Depending upon the constancy of the ratio of change between the variables, the correlation may be Linear or Non-linear Correlation.

**Linear Correlation**: If the amount of change in one variable bears a constant ratio to the amount of change in the other variable, then correlation is said to be linear. If such variables are plotted on a graph paper all the plotted points would fall on a straight line.

For example: If it is assumed that, to produce one unit of finished product we need 10 units of raw materials, then subsequently to produce 2 units of finished product we need double of the one unit.

| | | | | | | |
|---|---|---|---|---|---|---|
| Raw materials: X | 10 | 20 | 30 | 40 | 50 | 60 |
| Finished product: Y | 2 | 4 | 6 | 8 | 10 | 12 |

**Non-linear Correlation**: If the amount of change in one variable does not bear a constant ratio to the amount of change to the other variable, then correlation is said to be non-linear. If such variables are plotted on a graph, the points would fall on a curve and not on a straight line.

For example, if we double the amount of advertisement expenditure, then sales volume would not necessarily be doubled.

| | | | | | | |
|---|---|---|---|---|---|---|
| advertisement expenses: X | 10 | 20 | 30 | 40 | 50 | 60 |
| Sales Volume: Y | 2 | 6 | 4 | 8 | 12 | 10 |

---

**Problem1:** State in each case whether there is

(a) Positive Correlation (b) Negative Correlation $(c)$ No Correlation

| Sl No | Particulars | Solution |
|---|---|---|
| 1 | Price of commodity and its demand | Negative |
| 2 | Yield of crop and amount of rainfall | Positive |
| 3 | No of fruits eaten and hungry of a person | Negative |
| 4 | No of units produced and fixed cost per unit | Negative |
| 5 | No of girls in the class and marks of boys | No Correlation |
| 6 | Ages of Husbands and wife | Positive |
| 7 | Temperature and sale of woollen garments | Negative |
| 8 | Number of cows and milk produced | Positive |
| 9 | Weight of person and intelligence | No Correlation |
| 10 | Advertisement expenditure and sales volume | Positive |

**Methods of measurement of correlation:**

Quantification of the relationship between variables is very essential to take the benefit of study of correlation. For this, we find there are various methods of measurement of correlation, which can be represented as given below:

## Karl Pearson's Coefficient of Correlation

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

This method is also known as Product-Moment correlation coefficient

$$r = \frac{Cov(xy)}{\sigma_x \sigma_y} = \frac{\Sigma xy}{N\sigma_x \sigma_y} = \frac{\Sigma XY}{\sqrt{\Sigma X^2 \Sigma Y^2}} = \frac{\Sigma(x-\underline{X})(y-\underline{Y})}{\sqrt{\Sigma(x-\underline{X})^2 \Sigma(y-\underline{Y})^2}}$$

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

$\sigma_x$ =standard deviation of series $x$ and $\sigma_y$ = standard deviation of series $y$

## Properties

1. The Coefficient of correlation lies $b/w$ -1 & +1.

2. The Coefficient of correlations independent of change of origin & scale of measurements.

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

---

**Rank Correlation Coefficient**

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

formula: $\rho = 1 - 6\frac{\Sigma D^2}{N^3 - N}$

where: $\rho$ –Rank Coefficient of correlation

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

$N$-Number of paired observations.

---

**Properties**

1. The value of $\rho$ lies between $+1$ and $-1$.

2. If $\rho = 1$, then there is complete agreement in the order ofthe ranks & the direction of the rank is same.

3. If $\rho = -1$, then there is complete disagreement in the order of the ranks &they are in opposite directions.

---

**Equal or Repeated ranks**

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

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

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

$N$-Number of paired observations.

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

---

**Problem1:** From following information find the correlation coefficient between advertisement expenses and sales volume using Karl Pearson's coefficient of correlation method.

| Firm | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Advertisement expenses (In Laks) | 11 | 13 | 14 | 16 | 16 | 15 | 15 | 14 | 13 | 13 |
| Sales Volume (In Laks) | 50 | 50 | 55 | 60 | 65 | 65 | 65 | 60 | 60 | 50 |

---

**Solution:**

Let us assume that advertisement expenses are variable X and sales volume are variable Y.

Calculation of Karl Pearson's coefficient of correlation

| **Firm** | $X$ | **Deviation from mean** $X = x - \underline{x}$ | $X^2$ | $Y$ | **Deviation from mean** $Y = y - \underline{y}$ | $Y^2$ | **Product of $X$ and $Y$ series($XY$)** |
|---|---|---|---|---|---|---|---|
| **1** | **11** | **-3** | **9** | **50** | **-8** | **64** | **24** |
| **2** | **13** | **-1** | **1** | **50** | **-8** | **64** | **38** |
| **3** | **14** | **0** | **0** | **55** | **-3** | **9** | **0** |
| **4** | **16** | **2** | **4** | **60** | **2** | **4** | **4** |
| **5** | **16** | **2** | **4** | **65** | **7** | **49** | **14** |
| **6** | **15** | **1** | **1** | **65** | **7** | **49** | **7** |
| **7** | **15** | **1** | **1** | **65** | **7** | **49** | **7** |
| **8** | **14** | **0** | **0** | **60** | **2** | **4** | **0** |
| **9** | **13** | **-1** | **1** | **60** | **2** | **4** | **-2** |
| **10** | **13** | **-1** | **1** | **50** | **-8** | **64** | **8** |
| | $\sum X = 140$ | | $\sum X^2 = 22$ | $\sum Y = 580$ | | $\sum Y^2 = 360$ | $\sum XY = 70$ |

$$\underline{X} = \frac{\Sigma X}{n} = \frac{140}{10} = 14$$

$$\underline{Y} = \frac{\Sigma Y}{n} = \frac{580}{10} = 58$$

$$r = \frac{\Sigma XY}{\sqrt{\Sigma X^2 \Sigma Y^2}} = \frac{70}{\sqrt{22 \times 360}} = \frac{70}{88.9944} = 0.7866$$

**Interpretation:** From the above calculation it is very clear that there is high degree of positive correlation i.e. $r = 0.7866$, between the two variables. i.e. Increase in advertisement expenses leads to increased sales volume.

---

**Problem2:** Find Karl Pearson's coefficient of correlation from the following data.

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

**Solution:**

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

Coefficient of correlation $r = \frac{\Sigma XY}{\sqrt{\Sigma X^2 \Sigma Y^2}} = \frac{216}{\sqrt{(102)(472)}} = 0.98$

**Interpretation:** From the above calculation it is very clear that there is high degree of positive correlation i.e. $r = 0.98$, between the two variables. i.e. Height and Weights

**Problem3:** Calculate Coefficient of correlation for the following data.

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

**Solution:** Formula used: $r = \frac{\Sigma XY}{\sqrt{\Sigma X^2 \Sigma Y^2}}$

| **X** | **Y** | $X^2$ | $Y^2$ | **XY** |
|---|---|---|---|---|
| **12** | **14** | **144** | **196** | **168** |
| **9** | **8** | **81** | **64** | **72** |
| **8** | **6** | **64** | **36** | **48** |
| **10** | **9** | **100** | **81** | **90** |
| **11** | **11** | **121** | **121** | **121** |
| **13** | **12** | **169** | **144** | **156** |
| **7** | **3** | **49** | **9** | **21** |

| | | | | |
|---|---|---|---|---|
| $\sum X = 70$ | $\sum Y = 63$ | $\sum X^2 = 728$ | $\sum Y^2 = 651$ | $\sum XY = 676$ |

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

$$r = \frac{676}{\sqrt{(728)(651)}} = \frac{676}{688.42} = +0.98$$

**Interpretation:** From the above calculation it is very clear that there is high degree of positive correlation i.e. $r = 0.98$, between the two variables. i.e. X and Y

---

**Problem4:** Find the correlation coefficient between age and playing habits of the following students using Karl Pearson's coefficient of correlation method.

| Age | 15 | 16 | 17 | 18 | 19 | 20 |
|---|---|---|---|---|---|---|
| Number of Students | 250 | 200 | 150 | 120 | 100 | 80 |
| Regular Players | 200 | 150 | 90 | 48 | 30 | 12 |

---

**Solution:**

To find the correlation between age and playing habits of the students, we need to compute the percentages of students who are having the playing habit.

Percentage of playing habits $= \frac{No.of\ Regular\ Players}{Total\ No.of\ Students} \times 100$

Now, let us assume that ages of the students are variable X and percentages of playing habits are variable Y.

| Age $(x)$ | No of students | Regular Players | % of Playing Habits $(y)$ | **Deviation from mean** $X = x - \underline{x}$ | **Deviation from mean** $Y = y - \underline{y}$ | $X^2$ | $Y^2$ | **Product of $X$ and $Y$ series$(XY)$** |
|---|---|---|---|---|---|---|---|---|
| 15 | 250 | 200 | 80 | -2.5 | 30 | 6.25 | 900 | -75 |
| 16 | 200 | 150 | 75 | -1.5 | 25 | 2.25 | 625 | -37.5 |
| 17 | 150 | 90 | 60 | -0.5 | 10 | 0.25 | 100 | -5 |
| 18 | 120 | 48 | 40 | 0.5 | -10 | 0.25 | 100 | -5 |
| 19 | 100 | 30 | 30 | 1.5 | -20 | 2.25 | 400 | -30 |
| 20 | 80 | 12 | 15 | 2.5 | -35 | 6.25 | 1225 | -87.5 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| $\sum x = 105$ | | | $\sum y = 300$ | | | $\sum X^2 = 17.5$ | $\sum Y^2 = 3350$ | $\sum XY = -240$ |

$$\underline{X} = \frac{\Sigma X}{n} = \frac{105}{6} = 17.5$$

$$\underline{Y} = \frac{\Sigma Y}{n} = \frac{300}{6} = 50$$

$r = \frac{\Sigma XY}{\sqrt{\Sigma X^2 \Sigma Y^2}} = \frac{-240}{\sqrt{17.5 \times 3350}} = \frac{-2.40}{242.12} = -0.9912$

**Interpretation:** From the above calculation it is very clear that there is high degree of negative correlation i.e. $r = -O.9912$, between the two variables of age and playing habits. i.e. Playing habits among students decreases when their age increases.

---

**Problem5:** Given N = 10, $\sigma_x = 5.4, \sigma_y = 6.2$ and sum of product of deviation from the mean of X and Y is 66. Find the correlation coefficient.

---

**Solution:** given $N = 10$, $\sigma_x = 5.4$, $\sigma_y = 6.2$

$$\Sigma xy = \sum (x - \underline{x})(y - \underline{y}) = 66$$

$$r = \frac{\Sigma xy}{N\sigma_x\sigma_y} = \frac{66}{10 \times (5.4)(6.2)} = 0.1971$$

---

**Problem6:** Coefficient of correlation between X and Y is 0.3. Their covariance is 9. The variance of X is 16. Find the standard devotion of $Y$ series.

---

**Solution:** Given information:

$r = 0.3 \quad Cov(X, Y) = 9$ Var (X) $= 16$

$$r = \frac{Cov(X,Y)}{\sqrt{Var(X) \times Var(Y)}} \Rightarrow 0.3 = \frac{9}{\sqrt{16 \times Var(Y)}} \Rightarrow 0.3 = \frac{9}{4 \times \sqrt{Var(Y)}}$$

$0.3 \times 4 = \frac{9}{SD(Y)} \Rightarrow 1.2 = \frac{9}{SD(Y)} \Rightarrow$ SD $(Y) = \frac{9}{1.2} = 7.5$

Therefore, the standard deviation of $Y$ series $= \sigma(Y) = 7.5$

**Problem7:** Following are the rank obtained by 10 students in two subjects, statistics and mathematics. To what extent the knowledge of the students in two subjects is related

**Solution:**

| Ranks in statistics$(x)$ | Ranks in mathematics$(y)$ | $D = x - y$ | $D^2$ |
|---|---|---|---|
| 1 | 2 | -1 | 1 |
| 2 | 4 | -2 | 4 |
| 3 | 1 | 2 | 4 |
| 4 | 5 | -1 | 1 |
| 5 | 3 | 2 | 4 |
| 6 | 9 | -3 | 9 |
| 7 | 7 | 0 | 0 |
| 8 | 10 | -2 | 4 |
| 9 | 6 | 3 | 9 |
| 10 | 8 | 2 | 4 |
| | | | $\sum D^2 = 40$ |

$$\rho = 1 - 6\frac{\Sigma D^2}{N^3 - N} = 1 - \frac{6 \times 40}{10^3 - 10} = 1 - \frac{240}{990} = 1 - 0.24 = 0.76$$

**Problem8:** The ranks of 16 students in mathematics and statistics are as follows.

(1,1) (2,10) (3,3) (4,4) (5,5) (6,7) (7,2) (8,6) (9,8) (10,11) (11,15) (12,9) (13,14) (15,16) (16,13)

Calculate the rank correlation coefficient for proficiencies of this group in mathematics and statistics.

**Solution:**

| Ranks in statistics$(x)$ | Ranks in mathematics$(y)$ | $D = x - y$ | $D^2$ |
|---|---|---|---|
| 1 | 1 | 0 | 0 |
| 2 | 10 | -8 | 64 |
| 3 | 3 | 0 | 0 |
| 4 | 4 | 0 | 0 |
| 5 | 5 | 0 | 0 |
| 6 | 7 | -1 | 1 |
| 7 | 2 | 5 | 25 |

| | | | |
|---|---|---|---|
| **8** | **6** | **2** | **4** |
| **9** | **8** | **1** | **1** |
| **10** | **11** | **-1** | **1** |
| **11** | **15** | **-4** | **16** |
| **12** | **9** | **3** | **9** |
| **13** | **14** | **-1** | **1** |
| **14** | **12** | **2** | **4** |
| **15** | **16** | **-1** | **1** |
| **16** | **13** | **3** | **9** |
| | | | $\sum D^2 = 136$ |

$$\rho = 1 - 6\frac{\Sigma D^2}{N^3 - N} = 1 - \frac{6 \times 136}{16^3 - 16} = 1 - \frac{816}{4080} = 1 - 0.2 = 0.8$$

---

**Problem9:** Ten competitors in a musical test were ranked by the three judges A, B, and C in the following order.

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Ranks by A | 1 | 6 | 5 | 10 | 3 | 2 | 4 | 9 | 7 | 8 |
| Ranks by B | 3 | 5 | 8 | 4 | 7 | 10 | 2 | 1 | 6 | 9 |
| Ranks by C | 6 | 4 | 9 | 8 | 1 | 2 | 3 | 10 | 5 | 7 |

Using rank correlation method, discuss which pair of judges has the nearest approach to common likings in music.

---

**Solution: Here N = 10**

| **Ranks by B**$(x)$ | **Ranks by B**$(y)$ | **Ranks by C**$(z)$ | $D_1 = x - y$ | $D_2 = x - z$ | $D_3 = y - z$ | $D_1^2$ | $D_2^2$ | $D_3^2$ |
|---|---|---|---|---|---|---|---|---|
| **1** | **3** | **6** | **-2** | **-5** | **-3** | **4** | **25** | **9** |
| **6** | **5** | **4** | **1** | **2** | **1** | **1** | **4** | **1** |
| **5** | **8** | **9** | **-3** | **-4** | **-1** | **9** | **16** | **1** |
| **10** | **4** | **8** | **6** | **2** | **-4** | **36** | **4** | **16** |
| **3** | **7** | **1** | **-4** | **2** | **6** | **16** | **4** | **36** |
| **2** | **10** | **2** | **-8** | **0** | **8** | **64** | **0** | **64** |
| **4** | **2** | **3** | **2** | **1** | **-1** | **4** | **1** | **1** |
| **9** | **1** | **10** | **8** | **-1** | **-9** | **64** | **1** | **81** |
| **7** | **6** | **5** | **1** | **2** | **1** | **1** | **4** | **1** |
| **8** | **9** | **7** | **-1** | **1** | **2** | **1** | **1** | **4** |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | | | $\sum D_1^2 = 200$ | $\sum D_2^2 = 60$ | $\sum D_3^2 = 214$ |

$$\rho_1 = (x, y) = 1 - 6\frac{\Sigma D_1^2}{N^3 - N} = 1 - \frac{6 \times 200}{10^3 - 10} = 1 - \frac{40}{33} = -\frac{7}{33}$$

$$\rho_2 = (x, z) = 1 - 6\frac{\Sigma D_2^2}{N^3 - N} = 1 - \frac{6 \times 60}{10^3 - 10} = 1 - \frac{4}{11} = \frac{7}{11}$$

$$\rho_3 = (y, z) = 1 - 6\frac{\Sigma D_3^2}{N^3 - N} = 1 - \frac{6 \times 214}{10^3 - 10} = 1 - \frac{214}{165} = -\frac{49}{165}$$

Since $\rho_2 = (x, z)$ is maximum, we conclude that the pair of judges A and C has the nearest approach to common liking in music.

---

**Problem10:** A sample of 12 fathers and their elder sons gave the following data. Calculate the rank correlation coefficient.

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

---

**Solution:**

| **Fathers(X)** | **Sons(Y)** | **Rank(X)** | **Rank(Y)** | $d_i = x_i - y_i$ | ${d_i}^2$ |
|---|---|---|---|---|---|
| **65** | **68** | **9** | **5.5** | **3.5** | **12.25** |
| **63** | **66** | **11** | **9.5** | **1.5** | **2.25** |
| **67** | **68** | **6.5** | **5.5** | **1** | **1** |
| **64** | **65** | **10** | **11.5** | **-1.5** | **2.25** |
| **68** | **69** | **4.5** | **3** | **1.5** | **2.25** |
| **62** | **66** | **12** | **9.5** | **2.5** | **6.25** |
| **70** | **68** | **2** | **5.5** | **-3.5** | **12.25** |
| **66** | **65** | **8** | **11.5** | **3.5** | **12.25** |
| **68** | **71** | **4.5** | **1** | **-3.5** | **12.25** |
| **67** | **67** | **6.5** | **8** | **-1.5** | **2.25** |
| **69** | **68** | **3** | **5.5** | **-2.5** | **6.25** |
| **71** | **70** | **1** | **2** | **-1** | **1** |
| | | | | | $\sum {d_i}^2 = 72.5$ |

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

In $Y$: 68 appears 4 times, 66 appears twice & 65 appears twice.

Here $N = 12$.

**[In X series 68 appears 2 times so rank =** $\frac{4+5}{2} = 4.5$

**Next value is 67 also appears 2 times, so rank =** $\frac{6+7}{2} = 6.5$

**Next value is 66, then the next rank = 8**

**Similarly, in Y series 68 appears 4 times so the rank =** $\frac{4+5+6+7}{4} = 5.5$

**Next value is 67 so next rank = 8**

**Also 66 and 65 appears 2 times so their ranks are** $\frac{9+10}{2} = 9.5$ **and** $\frac{11+12}{2} = 11.5$**]**

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

$$= 1 - 6\left\{\frac{72.5 + \frac{1}{12}(4^3 - 4) + \frac{1}{12}(2^3 - 2) + \frac{1}{12}(2^3 - 2) + \frac{1}{12}(2^3 - 2) + \frac{1}{12}(2^3 - 2)}{N^3 - N}\right\}$$

$$1 - \frac{6(72.5 + 7)}{12(12^2 - 1)} = 0.722$$

---

**Problem11:** From the following data calculate the rank correlation coefficient after making adjustment for tied ranks

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| X | 48 | 33 | 40 | 9 | 16 | 16 | 65 | 24 | 16 | 57 |
| Y | 13 | 13 | 24 | 6 | 15 | 4 | 20 | 9 | 6 | 19 |

---

**Solution:** First we have to assign ranks to the variables. In this problem we are assigning lower value as first rank and highest value as last rank.

Here in X, 16 is repeated 3 times so rank = $\frac{2+3+4}{3} = 3$

Similarly, in Y, 6 is repeated 2 times so their rakes are $\frac{2+3}{2} = 2.5$ and the next value is 9 so for 9 the rank is 4

Again 13 is repeated 2 times so the next ranks are $\frac{5+6}{2} = 5.5$

| X | Ranks $(x)$ | Y | Ranks$(y)$ | $D = x - y$ | $D^2$ |
|---|---|---|---|---|---|
| 48 | 8 | 13 | 5.5 | 2.5 | 6.25 |
| 33 | 6 | 13 | 5.5 | 0.5 | 0.25 |
| 40 | 7 | 24 | 10 | -3 | 9 |
| 9 | 1 | 6 | 2.5 | -1.5 | 2.25 |
| 16 | 3 | 15 | 7 | 4 | 16 |
| 16 | 3 | 4 | 1 | 2 | 4 |
| 65 | 10 | 20 | 9 | 1 | 1 |
| 24 | 5 | 9 | 4 | 1 | 1 |
| 16 | 3 | 6 | 2.5 | 0.5 | 0.25 |
| 57 | 9 | 19 | 8 | 1 | 1 |
| | | | | | $\sum D^2 = 41$ |

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

$$1 - 6\left\{\frac{41 + \frac{1}{12}(3^3 - 3) + \frac{1}{12}(2^3 - 2) + \frac{1}{12}(2^3 - 2)}{10^3 - 10}\right\}$$

$$1 - \frac{6(41 + 3)}{990} = 0.733$$

---

**Regression equations**

Prediction or estimation of most likely values of one variable for specified values of the other is done by using suitable equations involving the two variables. Such equations are known as Regression Equations

---

**Regression equation of $y$ on $x$:**

$y - \underline{y} = b_{yx}(x - \underline{x})$ where $y$ is the dependent variable and $x$ is the independent variable and

$b_{yx}$ is given by

$$b_{yx} = \frac{\sum \quad (x-\underline{x})(y-\underline{y})}{\sum \quad (x-\underline{x})^2} \quad \textbf{OR} \quad b_{yx} = r\frac{\sigma_y}{\sigma_x} = \frac{\sum \quad XY}{\sum \quad x^2}$$

---

**Regression equation of $x$ on $y$:**

$x - \underline{x} = b_{xy}(y - \underline{y})$ where $y$ is the dependent variable and $x$ is the independent variable and

$b_{yx}$ is given by

$$b_{xy} = \frac{\sum \quad (x-\underline{x})(y-\underline{y})}{\sum \quad (y-\underline{y})^2} \quad \textbf{OR} \quad b_{xy} = r\frac{\sigma_x}{\sigma_y} = \frac{\sum \quad XY}{\sum \quad y^2}$$

$b_{yx}$ and $b_{xy}$ are called as regression coefficients of y on $x$ and $x$ on $y$ respectively.

Relation between correlation and regression coefficients:

$b_{yx} = r\frac{\sigma_y}{\sigma_X}$ and $b_{xy} = r\frac{\sigma_x}{\sigma_y}$

$b_{yx} \,.\, b_{xy} = r\frac{\sigma_y}{\sigma_X} \,.\, r\frac{\sigma_x}{\sigma_y} = r^2$

Hence $r = \pm\sqrt{b_{yx}b_{xy}}$

---

**Note:** In the above expression the component inside the square root is valid only when $b_{yx}$ and $b_{xy}$ have the same sign. Therefore, the regression coefficients will have the same sign.

---

**Problem1:** In trying to evaluate the effectiveness of its advertising campaign a company compiled the following information. Calculate the regression line of sales on advertising.

| Year | 1980 | 1981 | 1982 | 1983 | 1984 | 1985 | 1986 | 1987 |
|---|---|---|---|---|---|---|---|---|
| Advertisement in 1000 Rs | 12 | 15 | 15 | 23 | 24 | 38 | 42 | 48 |
| Sales in laks | 5 | 5.6 | 5.8 | 7 | 7.2 | 8.8 | 9.2 | 9.5 |

---

**Solution:** Let $x$ be advertising amount and $y$ be the sales amount.

Here, $n = 8, \underline{x} = \frac{217}{8} = 27.1, \underline{y} = \frac{58.1}{8} = 7.26$

We know that Regression equation of y on $x$ is given by $y - \underline{y} = b_{yx}(x - \underline{x})$.....(1)

Where $b_{yx} = r\frac{\sigma_y}{\sigma_x} = \frac{\sum \quad XY}{\sum \quad x^2}$

| X | Y | $X^2$ | XY |
|---|---|---|---|
| **12** | **5** | **144** | **60** |
| **15** | **5.6** | **225** | **84** |
| **15** | **5.8** | **225** | **87** |
| **23** | **7** | **529** | **161** |
| **24** | **7.2** | **576** | **172.8** |
| **38** | **8.8** | **1444** | **334.4** |
| **42** | **9.2** | **1764** | **386.4** |
| **48** | **9.5** | **2304** | **456** |
| $\sum X = 217$ | $\sum Y = 58.1$ | $\sum X^2 = 7211$ | $\sum XY = 1741.6$ |

$\therefore b_{yx} = \frac{1741.6}{7211} = 0.241$

Substituting this value in the in Eq…(1) $y$ on $x$ equation, we get,

$$y - 7.26 = 0.241(x - 27.1)$$

$$y = 0.241x - 6.5311 + 7.26$$

$$= 0.241x + 0.7289$$

Therefore, the required equation of Sales on Advertisement is $y = 0.7289 + 0.241x$

---

**Problem2:** In a study of the effect of a dietary component on plasma lipid composition, the following ratios were obtained on a sample of experimental annuals

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Measure of dietary component: (X) | 1 | 5 | 3 | 2 | 1 | 1 | 7 | 3 |
| Measure of Plasma lipid levl: (Y) | 6 | 1 | 0 | 0 | 1 | 2 | 1 | 5 |

(i) Obtain the two regression lines and hence predict the ratio of plasma lipid level with 4 dietary components.

(ii) Find the correlation coefficient between X and $Y$

---

**Solution:**

**(1)**

| X | Y | $X^2$ | $Y^2$ | XY |
|---|---|---|---|---|

| | | | | |
|---|---|---|---|---|
| **1** | **6** | **1** | **36** | **6** |
| **5** | **1** | **25** | **1** | **5** |
| **3** | **0** | **9** | **0** | **0** |
| **2** | **0** | **4** | **0** | **0** |
| **1** | **1** | **1** | **1** | **1** |
| **1** | **2** | **1** | **4** | **2** |
| **7** | **1** | **49** | **1** | **7** |
| **3** | **5** | **9** | **25** | **15** |
| $\sum X =$ **23** | $\sum Y =$ **16** | $\sum X^2 =$ **99** | $\sum Y^2 =$ **68** | $\sum XY =$ **36** |

Here $n = 8;\ \underline{x} = 2.875;\ \underline{y} = 2$

The Regression equation of $y$ on $x$ is given by $y - \underline{y} = b_{yx}(x - \underline{x})$

Where,

$$b_{yx} = r\frac{\sigma_y}{\sigma_x} = \frac{\sum XY}{\sum x^2}$$

$$b_{yx} = \frac{36}{99} = 0.3636$$

Hence the regression equation of y on $x$ is

$$y - 2 = 0.3636(x - 2.875)$$

$$= 0.3636x - 1.04535$$

$$y = 0.3636x - 1.04535 + 2$$

(i.e) $y = 0.95465 + 0.3636x$

when $x = 4$ (measure of dietary component) the plasmid lipid level is

$$y = 0.95465 + 0.3636(4)$$

$\Rightarrow y = 2.40905$

The Regression equation of $x$ on $y$ is given by $x - \underline{x} = b_{xy}(y - \underline{y})$ where,

$$b_{xy} = r\frac{\sigma_x}{\sigma_y} = \frac{\sum XY}{\sum y^2}$$

$$b_{xy} = \frac{36}{68} = 0.5294$$

Hence the regression equation of x on $y$ is

$$x - 2.875 = 0.5294(y - 2)$$

$$x = 0.5294y - 1.0588 + 2.875$$

$$x = 1.8162 + 0.5294y$$

**(ii)** The correlation coefficient between $x$ and $y$ is given by

$$r = \pm\sqrt{b_{yx}b_{xy}}$$

$$r = \pm\sqrt{0.3636 \times 0.5294} = \pm 0.4386$$

---

**Problem3:** From the data given below find (i) two regression lines (ii) coefficient of correlation between marks in Physics and marks in Chemistry (iii) most likely marks in Chemistry when marks in Physics is 78 (iv) most likely marks in Physics when marks in Chemistry is 92

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Marks in Physics:(X) | 72 | 85 | 91 | 85 | 91 | 89 | 84 | 87 | 75 | 77 |
| Marks in chemistry:(Y) | 76 | 92 | 93 | 91 | 93 | 95 | 88 | 91 | 80 | 81 |

---

**Solution:**

| **X** | **Y** | $X^2$ | $Y^2$ | **XY** |
|---|---|---|---|---|
| **72** | **76** | **5184** | **5776** | **5472** |
| **85** | **92** | **7225** | **8464** | **7820** |
| **91** | **93** | **8281** | **8649** | **8463** |
| **85** | **91** | **7225** | **8281** | **7735** |
| **91** | **93** | **8281** | **8649** | **8463** |
| **89** | **95** | **7921** | **9025** | **8455** |
| **84** | **88** | **7056** | **7744** | **7392** |
| **87** | **91** | **7569** | **8281** | **7917** |
| **75** | **80** | **5625** | **6400** | **6000** |
| **77** | **81** | **5929** | **6561** | **6237** |
| $\sum X = 836$ | $\sum Y = 880$ | $\sum X^2 = 70296$ | $\sum Y^2 = 77830$ | $\sum XY = 73954$ |

Here $n = 10$ $\underline{x} = 83.6$ $\underline{y} = 88$

The Regression equation of $y$ on $x$ is given by $y - \underline{y} = b_{yx}(x - \underline{x})$ where,

$$b_{yx} = r\frac{\sigma_y}{\sigma_x} = \frac{\sum \quad XY}{\sum \quad x^2}$$

$$b_{yx} = \frac{73954}{70296} = 1.052$$

Hence the regression equation of y on $x$ is

$$y - 88 = 1.052(x - 83.6)$$

$$= 1.052x - 87.94$$

$$\Rightarrow \; y = 0.06 + 1.052x$$

The Regression equation of $x$ on $y$ is given by $x - \underline{x} = b_{xy}(y - \underline{y})$

$$b_{xy} = r\frac{\sigma_x}{\sigma_y} = \frac{\sum \quad XY}{\sum \quad y^2}$$

$$b_{xy} = \frac{73954}{77830} = 0.950$$

Hence the regression equation of x on $y$ is

$$x - 83.6 = 0.950(y - 88)$$

$$= 0.950x - 83.6$$

$$\Rightarrow \; x = 0 + 0.950y$$

(ii) The correlation coefficient between $x$ and $y$ is given by

$$r = \pm\sqrt{b_{yx}b_{xy}}$$

$$r = \pm\sqrt{1.052 \times 0.950} = \pm 0.9996$$

(iii) To find the most likely marks in Chemistry when marks in Physics is 78, we must use

the regression equation of $y$ on $x$ given by $y = 0.06 + 1.052x$, Substituting the value of x as 78 in the above equation, we get,

$$y = 0.06 + 1.052(78) = 82.116$$

Hence the marks in Chemistry is 82.116

(iv) To find the most likely marks in Physics when marks in Chemistry is 92, we must use

the regression equation of x on $y$ given by $x = 0 + 0.950y$

Substituting the value of y as 92 in the above equation, we get,

$$x = 0 + 0.950(92)$$

$$x = 87.4$$

Hence the marks in Physics is 87.4

---

**Problem4:** For a given series of values, the following data were obtained,

$\underline{x} = 36, \underline{y} = 85, \sigma_x = 11, \sigma_y = 8$ and $r = 0.66$.

Find (i) two regression equations (ii) estimation of $x$ when $y = 75$.

---

**Solution:**

We have $b_{yx} = r\frac{\sigma_y}{\sigma_x} = 0.66 \times \frac{8}{11} = 0.48$

and $b_{xy} = r\frac{\sigma_x}{\sigma_y} = 0.66 \times \frac{11}{8} = 0.9075$

(i) The Regression equation of $y$ on $x$ is given by

$$y - \underline{y} = b_{yx}(x - \underline{x})$$

$\Rightarrow y - 85 = 0.48(x - 36)$

$\Rightarrow y = 67.28 + 0.48x$

The Regression equation of $x$ on $y$ is given by

$$x - \underline{x} = b_{xy}(y - \underline{y})$$

$\Rightarrow x - 36 = 0.9075(y - 85)$

(i.e.) $x = -41.13 + 0.9075y$

(ii) To estimate the value of x when $y = 75$, we use the regression line ofx on $y$

$x = -41.13 + 0.9075y$ Substituting $y = 75, x = -41.13 + 0.9075(75)$

Therefore $x = 26.93$

---

**Problem5:** For a certain X and $Y$ series which are correlated, the regression lines are

$8x - 10y = -66$ and $40x - 18y = 214$

Find (i) the correlation coefficient between them and (ii) the mean of the two series.

**Solution:** The given regression equations are

$8x - 10y = -66 \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (1)$ and $40x - 18y = 214 \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots. (2)$

Let us suppose that the equation (1) is the equation of line of regression of $y$ on $x$ and (2) as the equation of the line of regression of x on $y$, after rewriting (1) and (2), we get

$y = \frac{66}{10} + \frac{8}{10}x$ which gives the value of $b_{yx} = \frac{8}{10}$

$x = \frac{214}{40} + \frac{18}{40}y$ which gives the value of $b_{xy} = \frac{18}{40}$

Now $r = \pm\sqrt{b_{yx}b_{xy}} = \pm\sqrt{\frac{8}{10} \times \frac{18}{40}} = \pm 0.6$

(ii) Since both the line of regression passes through the mean values $\underline{x}$ and $\underline{y}$, the point $(\underline{x}, \underline{y})$ must satisfy the given two regression lines.

Therefore, 8 $\underline{x} - 10\underline{y} = -66$, 40 $\underline{x} - 18\underline{y} = 214$

Solving the above two equations we get $\underline{x} = 13$ and $\underline{y} = 17$

**Important Note:** In the above problem in part (i), if we take equation (1) as the line of regression of $x$ on $y$, we get, $x = -\frac{66}{8} + \frac{10}{8}y$, and hence $b_{xy} = \frac{10}{8}$ and if we take equation (2) as the line of regression of y on $x$, we get, $y = -\frac{214}{18} + \frac{40}{18}x$ and hence $b_{yx} = \frac{40}{18}$

Therefore, $r = \pm\sqrt{b_{yx}b_{xy}} = \pm\sqrt{\frac{10}{8} \times \frac{40}{18}} = \pm 1.67$

But the value of $r$ cannot exceed unity. Hence the assumptions that line (1) is line of regression of x on $y$ and the line (2) is line of regression of $y$ on $x$ are wrong.

---

**FITTING OF CURVES**

1 Principle of Least Squares

2 Fitting of Straight Line

3 Fitting of Second-Degree Parabola

4 Fitting of a Power Curve $y = aX^b$

5 Fitting of Exponential Curve $y = ab^x$

6 Fitting of Exponential Curve $Y = ae^{bX}$

---

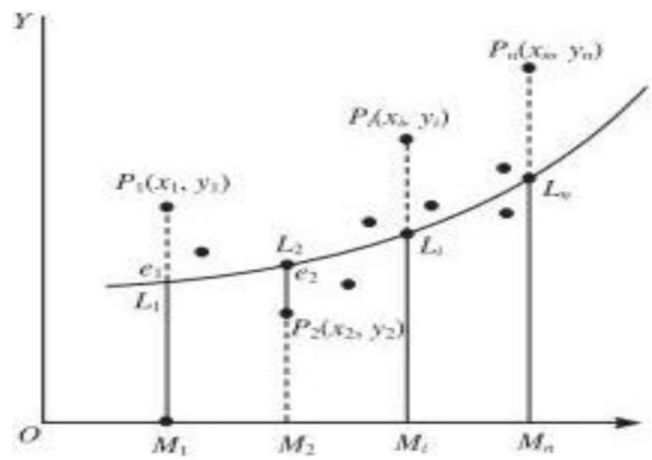
**Introduction**: All the methods that you have learnt earlier of this course were based on the uni-variate distributions, i.e. all measures analyzed only single variable. But many times, we have data for two or more variables and our interest is to know the best functional relationship between variables that given data describes. In this unit, you will learn how to fit the various functions such as straight line, parabola of the second degree, power curve and exponential curves. Curve fitting has theoretical importance in regression and correlation analysis while practically it is used to present the relationship by simple algebraic expression. All these methods can be used to estimate the values of the dependent variable for the specific value of the independent variable.

---

**PRINCIPLE OF LEAST SQUARES**

**Principle of Least Squares**

The principle of least squares is one of the most popular methods for finding the curve of best fit to a given data set $(x_i, y_i)$, $i = 1,2,3, n$.




Let $y = f(x)$ be the equation of the curve to be fitted to the given set of points $P_1(x_1, y_1)$ ,

$P_2(x_{2t}y_2), \cdots, P_n(x_{n/}y_n)$. Then at a point $x = x_j$, the observed value of the ordinate is $P_jM_j = y_j$ say and let the expected (theoretical) value be $f(x_j)$ , shown by $L_iM_i$ in the adjoining figure. The difference between the observed and expected values is the error $e_i = P_iL_i$

Then $e_1 = y_1 - f(x_1)$, $e_2 = y_2 - f(x_2)$, …… $e_n = y_n - f(x_n)$

Squaring each error $e_i$ (to take care of negative errors) and adding, we get

$E = e_1^2 + e_2^2 + \dots + e_n^2 = \sum e_i^2 = \sum (y_i - f(x_i))^2$ The curve of best fit is that for which $E$ is minimum. This is called the principle of least squares.

---

**FITTING OF STRAIGHT** $LINE$: $y = a + bX$

The normal equation is can be written as

$$\sum y = na + b\sum x \dots\dots (1)$$

$$\sum yx = a\sum x + b\sum x^2 \dots\dots (2)$$

Values of a and $b$ are obtained by solving equations (1) and (2). With this value of a and $b$, straight line $y = a + bX$ is the line of best fit to the given set of points $(x_i, y_i)$

where i= $1,2, \dots, n$.

---

**Problem1:** Fit a straight line to the following data.

| | | | | | | |
|---|---|---|---|---|---|---|
| X | 1 | 6 | 11 | 16 | 20 | 26 |
| Y | 13 | 16 | 17 | 23 | 24 | 31 |

---

**Solution:** Let the straight line be $y = a + bX$ and to obtain a and $b$ for this straight line the normal equations are

$\sum y = na + b\sum x$ and $\sum xy = a\sum x + b\sum x^2$

Here, there is need $of \sum y, \sum x, \sum xy$ and $\sum x^2$ which are obtained by the following table

| X | Y | $X^2$ | XY |
|---|---|---|---|
| 1 | 13 | 1 | 13 |
| 6 | 16 | 36 | 96 |
| 11 | 17 | 121 | 187 |
| 16 | 23 | 256 | 368 |
| 20 | 24 | 400 | 480 |
| 26 | 31 | 676 | 806 |
| $\sum X = 80$ | $\sum Y = 124$ | $\sum X^2 = 1490$ | $\sum xy = 1950$ |

Substituting the values of $\sum y, \sum x, \sum xy$ and $\sum x^2$ in the normal equations, we get

$$124 = 6a + 80b \ldots\ldots\ldots (1)$$

$$1950 = 80a + 1490b \ldots . (2)$$

Now we solve equations (1) and (2).

Multiplying equation (1) by 80 and equation (2) by 6, $i.e.$

$$124 = 6a + 80b\ ] \times 80 \text{ and } 1950 = 80a + 1490b] \times 6$$

we get,

$$9920 = 480a + 6400b \ldots\ldots..(3)$$

$$11700 = 480a + 8940b \ldots\ldots\ldots.(4)$$

Subtracting (3) from (4), we obtain

$$1780 = 2540b \Rightarrow b = 1780/2540 = 0.7008$$

Substituting the value of b in equation (1), we get

$$124 = 6a + 80 \times 0.7008$$

$$124 = 6a + 56.064$$

$$67.936 = 6a \Rightarrow a = 11.3227$$

with these values of a and $b$ the line ofbest fit is $y = 11.3227 + 0.7008$ X.

Now let us do one problem for fitting of straight line.

---

**Problem2:** Fit a straight line to the following data.

| | | | | | |
|---|---|---|---|---|---|
| X | 6 | 7 | 8 | 9 | 11 |
| Y | 5 | 4 | 3 | 2 | 1 |

**Ans:** $y = 9.6486 - 0.8108$

---

**Problem3:** By the method of least squares, find a straight line that best fits the following data points.

| | | | | | |
|---|---|---|---|---|---|
| X | 0 | 1 | 2 | 3 | 4 |
| Y | 1 | 2.9 | 4.8 | 6.7 | 8.6 |

**Ans:** $y = 1.9x + 1$

---

## FITTING OF SECOND-DEGREE PARABOLA

Let $Y = a + bX + cX^2 \ldots\ldots\ldots\ldots (1)$

be the second-degree parabola and we have a set of n points $(x_i, y_i)$ ; $i = 1,2, \ldots, n$. Here the problem is to determine $a$, $b$ and $c$ so that the equation of second-degree parabola given in equation (1) is the best fit equation of parabola

The normal equation is can be written as

$\sum y = na + b\sum x + c\sum x^2$……(1)

$\sum xy = a\sum x + b\sum x^2 + c\sum x^3$…..(2)

$\sum x^2y = a\sum x^2 + b\sum x^3 + c\sum x^4$…..(3)

Values of $a$, $b$ and $c$ are obtained by solving equations (1), (2) and (3). With these values of $a$, $b$ and $c$, the second-degree parabola $y = a + bX + cX^2$ is the best fit.

---

**Problem1:** Fit a second-degree parabola for the following data:

| X | 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| Y | 1 | 3 | 4 | 5 | 6 |

---

**Solution**: Let $y = a + bX + cX^2$ be the second-degree parabola and we have to

determine $a$, $b$ and $c$. Normal equations for second degree parabola are

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

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

and

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

To solve above normal equations, we need $\sum y, \sum x, \sum xy, \sum x^2y, \sum x^2,$ $\sum x^3, \sum x^4$

which are obtained from following table:

| $x$ | $y$ | $xy$ | $x^2$ | $x^2y$ | $x^3$ | $x^4$ |
|---|---|---|---|---|---|---|

| | | | | | | |
|---|---|---|---|---|---|---|
| 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 1 | 3 | 3 | 1 | 3 | 1 | 1 |
| 2 | 4 | 8 | 4 | 16 | 8 | 16 |
| 3 | 5 | 15 | 9 | 45 | 27 | 81 |
| 4 | 6 | 24 | 16 | 96 | 64 | 256 |
| $\sum x = 10$ | $\sum y = 19$ | $\sum xy = 50$ | $\sum x^2 = 30$ | $\sum x^2y = 160$ | $\sum x^3 = 100$ | $\sum x^4 = 354$ |

Substituting the values of $\sum y, \sum x, \sum xy, \sum x^2y, \sum x^2, \sum x^3, \sum x^4$ in above normal equations, we have

$$19 = 5a + 10b + 30c \ldots\ldots(1)$$

$$50 = 10a + 30b + 100c \ldots\ldots\ldots..(2)$$

$$160 = 30a + 100b + 354c \ldots\ldots(3)$$

Now, we solve equations (1), (2) and (3).

Multiplying equation (1) by 2, we get

$$38 = 10a + 20b + 60c \ldots..(4)$$

Subtracting equation (4) from equation (2)

$$50 = 10a + 30b + 100c$$

$$\underline{38 = 10a + 20b + 60c}$$

$$12 = 10b + 40c \ldots\ldots\ldots\ldots(5)$$

Multiplying equation (2) by 3, we get

$$150 = 30a + 90b + 300c \ldots\ldots\ldots..(6)$$

Subtracting equation (6) from equation (3), we get

$$160 = 30a + 100b + 354c$$

$$\underline{150 = 30a + 90b + 300c}$$

$$10 = 10b + 54c\text{......(7)}$$

Now we solve equation (5) and (7)

Subtracting equation (27) from equation (29), we get

$$10 = 10b + 54c$$

$$\underline{12 = 10b + 40c}$$

$$-2 = 14c$$

$c = -2/14 \Rightarrow c = -0.1429$

Substituting the value of $c$ in equation (7), we get

$$10 = 10b + 54 \times (-0.1429)$$

$$10 = 10b - 7.7166$$

$$17.7166 = 10b$$

$$b = 1.7717$$

Substituting the value of b and $c$ in equation (1), we get

$$19 = 5a + 10 \times (1.7717) + (-0.1429 \times 30)$$

$$19 = 5a + 17.717 - 4.287$$

$$a = 1.114$$

Thus, the second degree of parabola of best fit is

$$Y = 1.114 + 1.7717x - O.1429\, x^2$$

---

**Problem2:** Fit a parabola $y = ax^2 + bx + c$ to the given data

| | | | | | |
|---|---|---|---|---|---|
| $x$ | 10 | 12 | 15 | 23 | 20 |
| $y$ | 14 | 17 | 23 | 25 | 21 |

**ANS:** $y = -0.07x^2 + 3.01x - 8.73$

---

**Problem3:** Fit a $2^{nd}$ parabola to the given data

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

**ANS:** $y = 0.009x^2 + 0.77x + 0.195$

---

## FITTING OF A POWER CURVE $Y = a\,X^b$

Let the curve be $Y = ax^b$ ……………….(1)

Taking log on both sides

$\log Y = \log(aX^b)$

$= \log a + b \log X$

Let $U = \log Y$; $A = \log a$ and $V = \log X$

Then we get $U = A + bV$

Which is in the form of straight line in U and V

Hence the normal equations are

$$\sum U = nA + b\sum V$$

$$\sum UV = A\sum V + b\sum V^2$$

Solving these normal equations we get the values of A and b. Consequently, we get

$a = \text{Anti}\log(A)$.

Substituting these values in (1). We get the required power curve of the form $Y = ax^b$

---

**Problem1:** Fit the power curve of the form $Y = ax^b$ for the following data

| X | 1 | 2 | 4 | 6 |
|---|---|---|---|---|
| Y | 6 | 4 | 2 | 2 |

---

**Solution:** Let the curve be $Y = ax^b$ ……………….(1)

Taking log on both sides

LogY = log($aX^b$)

= loga + b logX

Let U = logY; A= log a and V = logX

Then we get U= A + b V

Which is in the form of straight line in U and V

Hence the normal equations are

$$\sum U = nA + b\sum V$$

$$\sum UV = A\sum V + b\sum V^2$$

| X | Y | V=logX | U= logY | $V^2$ | UV |
|---|---|---|---|---|---|
| 1 | 6 | 0 | 1.79 | 0 | 0 |
| 2 | 4 | 0.693 | 1.386 | 0.48 | 0.96 |
| 4 | 2 | 1.386 | 0.693 | 1.92 | 0.96 |
| 6 | 2 | 1.79 | 0.693 | 3.2 | 1.24 |
| TOTAL | | $\sum V =$ 3.87 | $\sum U =$ 4.56 | $\sum V^2 =$ 5.6 | $\sum UV$ =3.16 |

**Here n= 4 and substitute the above values in normal equations we have,**

$$4.56 = 4A + 3.87b$$

$$3.16 = 3.87\,A + 5.6b$$

Solving these equations, we get A=1.786; b= -0.672

And a= Anti log (1.786) = 5.965

Hence the required power curve is Y = 5.965 $x^{-0.672}$

---

**Problem2:** _Fit the power 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 | 6.10 | 6.80 | 7.50 |

---

**Problem3:** Fit the power curve of the form Y= $ax^b$ for the following data

| X | 1 | 3 | 5 | 7 | 9 |
|---|---|---|---|---|---|
| Y | 2 | 7 | 10 | 11 | 9 |

---

**<u>Fitting of an exponential curve of the form Y= a $e^{bX}$</u>**

Let the curve be Y= a $e^{bX}$ ………………(1)

Taking log on both sides

LogY = log($ae^{bX}$)

= loga + b X

Let U = log Y and A = log a

Then we get U= A + b X

Which is in the form of straight line in U and X

Hence the normal equations are

$$\sum U = nA + b\sum X$$

$$\sum UX = A\sum X + b\sum X^2$$

Solving these normal equations, we get the values of A and b. Consequently, we get a = Anti log (A) = $e^A$

Substituting these values in (1). We get the required power curve of the form Y= $ae^{b\,X}$

---

**Problem1:** Fit an exponential curve of the form Y= $ae^{b\,X}$ to the following data

| X | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|
| Y | 144 | 172.8 | 207.4 | 248.8 | 298.6 |

---

**Solution:** Let the curve be Y= a $e^{bX}$ ………………(1)

Taking log on both sides

LogY = log($ae^{bX}$)

= loga + b X

Let U = log Y and A = log a Then we get U= A + b X

Which is in the form of straight line in U and X

Hence the normal equations are

$$\sum U = nA + b\sum X$$

$$\sum UX = A\sum X + b\sum X^2$$

| X | Y | U=logY | UX | $X^2$ |
|---|---|---|---|---|
| 2 | 144 | 4.97 | 9.94 | 4 |
| 3 | 172.8 | 5.15 | 15.45 | 9 |
| 4 | 207.4 | 5.33 | 21.32 | 16 |
| 5 | 248.8 | 5.52 | 27.60 | 25 |
| 6 | 298.6 | 5.69 | 34.14 | 36 |
| $\sum X =$ 20 | | $\sum U =$26.66 | $\sum UX =$108.45 | $\sum X^2 =$ 90 |

**Here n= 5 and substitute the above values in normal equations we have,**

$$26.66 = 5A + 20b$$

$$108.45 = 20A + 90b$$

Here n= 5 **and substitute the above values in normal equations we have,** A = 4.608 and a= 100.28; b= 0.181

Hence the required curve is Y= 100.28 $e^{0.181X}$

---

**Problem2:** Fit an exponential curve of the form Y= $ae^{b\,X}$ to the following data

| X | 0 | 2 | 4 |
|---|---|---|---|
| Y | 5.012 | 10 | 31.62 |

---

**Fitting of an exponential curve of the form $Y= a\, b^X$**

Let the curve be $Y= a\, b^X$ ………………(1)

Taking log on both sides

$\log Y = \log(ab^X)$

$= \log a + X \log b$

Let $U = \log Y$; $A = \log a$ and $B= \log b$

Then we get $U= A + BX$

Which is in the form of straight line in U and X

Hence the normal equations are

$$\sum U = nA + B\sum X$$

$$\sum UX = A\sum X + B\sum X^2$$

Solving these normal equations, we get the values of A and B. Consequently, we get a = Anti log (A) = $e^A$ and

b = Anti log (B) = $e^B$.

Substituting these values in (1). We get the required power curve of the form $Y= ab^X$

---

**Problem1:** Fit an exponential curve of the form $Y= ab^X$ to the following data

| X | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|
| Y | 144 | 172.8 | 207.4 | 248.8 | 298.6 |

**Solution:** Let the curve be Y= a $b^X$ ……………….(1)

Taking log on both sides

LogY = log($ab^X$)

= loga + X log b

Let U = log Y; A = log a and B= log b

Then we get U= A + BX

Which is in the form of straight line in U and X

Hence the normal equations are

$$\sum U = nA + B\sum X$$

$$\sum UX = A\sum X + B\sum X^2$$

| X | Y | U=logY | UX | $X^2$ |
|---|---|---|---|---|
| 2<br>3<br>4<br>5<br>6 | 144<br>172.8<br>207.4<br>248.8<br>298.6 | 2.1584<br>2.2375<br>2.3168<br>2.3959<br>2.4751 | 4.3168<br>6.7125<br>9.2672<br>11.9795<br>14.8506 | 4<br>9<br>16<br>25<br>36 |
| $\sum X =$ 20 | | $\sum U =$11.5 837 | $\sum UX =$47.1 266 | $\sum X^2 =$ 90 |

Here n= 5 **and substitute the above values in normal equations we have,**

**11.5837 =5A +20B**

**47.1266 = 20A + 90B**

Solving these equations, we get A= 2.0001; B= 0.0792

a= 100 and b= 1.2

Hence the required curve is Y = 100 $(1.2)^X$

**Problem2:** Fit an exponential curve of the form Y= $ab^X$ to the following data

| | | | |
|---|---|---|---|
| X | 0 | 2 | 4 |
| Y | 5.012 | 10 | 31.62 |