Lecture Notes

BCS301 : Mathematics for Computer Science

Module 1 - Probability Distributions

Prepared by

Dr. Shantha Kumari. K

AJIET, Mangaluru

Email : shanthakk99@gmail.com

Table of Contents

Module 1

Probability Distributions

Syllabus : Review of basic probability theory, Random variables (discrete and continuous), probability mass and density functions, Mathematical expectation, mean and variance, Binomial, Poisson and normal distributions- problems (derivations for mean and standard deviation for Binomial and Poisson distributions only)-Illustrative examples. exponential distribution.

1.1 Review of basic probability theory :

- **Random Experiment :** An experiment is a random experiment if its outcome cannot be predicted precisely. One out of a number of outcomes is possible in a random experiment. A single performance of the random experiment is called a trial.
- **Outcome :** An outcome is a possible result of an experiment.
- **Sample Space :** The sample space S is the collection of all possible outcomes of a random experiment. The elements of S are called sample points. A sample space is called discrete if it is a finite or a countably infinite set. An uncountable sample space is called a continuous sample space.
- **Event :** An event is a subset of the sample space. i.e. a subset of possible outcomes for our experiment.

Events are denoted by $A, B, C \cdots$

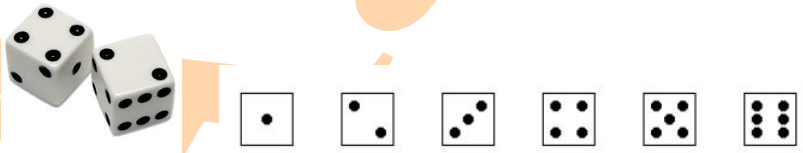
S is the **certain event** (sure to occur) and Φ is the **impossible event**.

Example 1 : Tossing a fair coin The possible outcomes are H (head) and T (tail). The associated sample space is $S = \{H, T\}$. It is a finite sample space. The events associated with the sample space S are: S, H, T and Φ.

Example 2 : Consider the experiment of tossing a coin twice.
Then $S = \{HH, HT, TH, TT\}$. For a small sample space like this, one can readily list all of the possible events. In this case, there are 16 possible events. Some of them are $A = \{HH\}, B = \{HT\}, C = \{TH\}, D = \{TT\}, E = \{HH, HT\}, F = \{HH, TH\}$ and so on.

Example 3 : Tossing a fair coin thrice Consider the experiment of flipping a coin three times and recording the possible outcomes of the three flips. In this case, the sample space is $S = \{TTT, TTH, THT, HTT, THH, HTH, HHT, HHH\}$ Here $\{HHH\}, \{HHT, THH, TTT\}, \{HHH, TTT\}$ etc. are some of the events.

Example 4 : Throwing a fair die The possible 6 outcomes are:



The associated finite sample space is $S = \{1, 2, 3, 4, 5, 6\}$. Some events are , A = The event of getting an odd face = $\{1, 3, 5\}$ B = The event of getting a six = $\{6\}$, and so on.

- **Equally Likely Events (Equiprobable events) :** Two or more events are equally likely if they have equal chance of occurrence. That is, equally likely events are

such that none of them has greater chance of occurrence than the others.

Example 1. While tossing a fair coin, the outcomes 'Head' and 'Tail' are equally likely.

Example 2. While throwing a fair die, the events $A = \{2, 4, 6\}, B = \{1, 3, 5\} \& C = \{1, 2, 3\}$ are equally likely.

- **Mutually Exclusive events (Disjoint events) :** Two or more events are mutually exclusive if only one of them can occur at a time. That is, the occurrence of any of these events totally excludes the occurrence of the other events. Mutually exclusive events cannot occur together.

Example 1. While tossing a coin, the outcomes 'Head' and 'Tail' are mutually exclusive because when the coin is tossed once, the result cannot be Head as well as Tail.

Example.2. While throwing a die, the events $A = \{2, 4, 6\}$, $B = \{3, 5\}$ and $C = \{1\}$ are mutually exclusive.

- **Mathematical definition of probability :** If the outcome of a trial consists n exhaustive, mutually exclusive, equally possible cases, of which m of them are favourable cases to an event E, then the probability of the happening of the event E, usually denoted by $P(E)$ or simply p is defined to be equal to $\frac{m}{n}$.

$$i.e.. \ P(E) = \frac{\text{(number of favourable cases)}}{\text{(number of possible cases)}} = \frac{m}{n}$$

- **Axioms of Probability :**

(i) **Axiom 1:** For any event A, $P(A) \geq 0$
(a negative probability does not make sense).

(ii) **Axiom 2 :** If S is the sample space for a given experiment, $P(S) = 1$ (i.e. probabilities are normalized so that the maximum value is unity).

(iii) **Axiom 3 :** If $A \cap B = \Phi$, then $P(A \cup B) = P(A) + P(B)$

Addition Theorem : For any sets A and B (not necessarily mutually exclusive),

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

The addition theorem of probability for three events is given by

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

The probability of an event A occurring when it is known that some event B has occurred, is called a conditional probability of the event given that B has occurred and denoted by $P(A|B)$).

$P(A|B)$, is defined by

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \text{ if } P(B) > 0$$

If $P(B) = 0, P(A|B)$ is not defined.

This formula offers a convenient way to compute joint probabilities:

$$P(A \cap B) = P(A)P(B|A) = P(B)P(A|B)$$

Two events A and B are independent (or statistically independent) if probability of occurrence of one event does not change the probability of occurrence of other event. (i.e. if $P(A|B) = P(A)$). Note that if $P(A|B) = P(A)$, then the following two conditions also hold.

$$P(B|A) = P(B)$$

and

$$P(A \cap B) = P(A)P(B)$$

1.2 Random Variables

A **random variable** is a function that associates a real number with each element in the sample space. In other words it is a function from the sample space S to the set of all real numbers; denoted as $\boldsymbol{X : S \mapsto \mathbb{R}}$

Example 1: While tossing a coin, suppose that the value 1 is associated for the outcome 'head' and 0 for the outcome 'tail'. In other words, let X= no. of heads.
We have the sample space $\boldsymbol{S = \{H, T\}}$ and if X is the random variable then $\boldsymbol{X(H) = 1}$ & $\boldsymbol{X(T) = 0}$.
Hence range of $\boldsymbol{X = \{0, 1\}}$

Example 2: For example, consider the random experiment of tossing three fair coins up.
Then S = {HHH, HHT, HTH, THH, TTH, THT, HTT, TTT}.
Define X as the number of heads that appear.
Hence, X (HHH) =3 , X (HHT) =2, X (HTH) =2, X (THH) =2 , X (HTT) =1, X(THT) =1 , X (TTH) =1 and X(TTT) =0.
Here the image set of the random variable may be written as $\boldsymbol{X = \{0, 1, 2, 3\}}$.

1.3 Discrete random variable

A **discrete random variable** is a random variable X whose possible values can be listed as a finite set of values or countably infinite set of values.

Example : In the Experiment of tossing a coin twice, if X= such as no. of heads, then the values taken by X are listed as $\boldsymbol{X = 0, 1, 2}$. Hence X is discrete random variable.

Exampe : Other Examples for discrete random variable are the size of a family, No. of accidents on road, No. of customers in a bank, No. of cars manufactured by the company, The no. of transactions in a bank per day etc

1.4 Continuous Random Variable

A random variable 'X' which takes all possible values in a given interval is called a **continuous random variable**.

Example : If X takes all values in a interval [0,1], then we can not list the values taken by X. Hence X is a continuous random variable.
Other examples are, The temperature at a location, The life time of electronic component, The reaction temperature at chemical laboratory, Current in a semi-conductor diode etc.

Note :
Remember that discrete random variables can take only a countable number of possible values. On the other hand, a continuous random variable $\boldsymbol{X}$ has a range in the form of an interval.

1.5 Probability mass function (pmf) :

Let $\boldsymbol{X}$ be a discrete random variable defined on the Sample space $\boldsymbol{S}$, and let the values of X are $\boldsymbol{x_1, x_2, x_3, \cdots, x_n}$. The function $\boldsymbol{p(x_i) = P(X = x_i)}$, that assigns a probability to each possible value of the random variable is said to be a **probability mass function(pmf)**(or probability distribution) if it satisfies the following properties.

(a) **Non-negativity** : $\boldsymbol{P(x_i) \geq 0}$, for all $\boldsymbol{x_i \in X}$

(b) **Normality:** $\Sigma_{i=1}^{\infty} P(x_i) = 1$ for all $x_i \in X$

Note : The probability that a random variable X takes a value in the closed interval $[a, b]$ is given by

$$P(a \leq X \leq b) = \sum_{a \leq x \leq b} p(x)$$

Note : Sometimes probability mass function is also called as probability density function

Example : Consider tossing of three coins We get
$S = \{HHH, HTT, THH, TTH, HHT, HTH, THT, TTT\}$.
If X = no. of heads. then we have the following possibilities.

X=Number of heads	0	1	2	3
Number of occurrence	1	3	3	1
$P(x)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

We can observe that $P(x_i) \geq 0$, for all $x_i \in X$ and $\Sigma_{i=1}^{\infty} P(x_i) = 1$
Hence $p(x)$ given above is a pmf.

1.6 Probability density function

Let X be a continuous random variable. Then a function f(x) defined for all values of X is said to be a **probability density function(pdf)** if it satisfies the following properties.

(a) non-negativity i.e., $f(x) \geq 0$, for all x.

(b) Normality : i.e. $\int_{-\infty}^{\infty} f(x)dx = 1$ (i.e. The area under a pdf curve is always 1).

Note : The probability that a random variable X takes a value in the (open or closed) interval $[a, b]$ is given by the integral of a function. i.e.

$$P(a \leq X \leq b) = \int_a^b f(x)dx$$

1.7 Mean and Variance

Mean and Variance of discrete random variable :

- If X is a **discrete** random variable with pmf $p(x)$, then **expected value** of X is defined as
  $E(X) = \sum_x xp(x).$
  This is also sometimes called the **mean** of the random variable X and denoted as $\bar{x}$ or μ.

- The expected value of X^2 is given by
  $E(X^2) = \sum_x x^2 p(x).$
- The quantity $E(X-\mu)^2 = E(X^2) - (\mu)^2$ is called the **variance** of the random variable X and is denoted Var(X) or $V(X)$ or σ^2.
- The square root of the variance, $\sigma \equiv \sqrt{Var(X)}$ is called the **standard deviation**.

Mean and Variance of continuous random variable :

- If X is a **continuous** random variable with pdf $f(x)$, then the expected value of X is defined by
  $E(X) = \int_{-\infty}^{\infty} x\, f(x)\, dx$
  The expected value of X^2 is
  $E(X^2) = \int_{-\infty}^{\infty} x^2\, f(x)\, dx,$
- The quantity $E(X-\mu)^2 = E(X^2) - (\mu)^2$ is called the **variance** of the random variable X and is denoted Var(X) or $V(X)$ or σ^2.
- The square root of the variance, $\sigma \equiv \sqrt{Var(X)}$ is called the **standard deviation**.

Problem 1. *The probability density function of a random variable X is*

X	*0*	*1*	*2*	*3*	*4*	*5*	*6*
p(X)	*k*	*3k*	*5k*	*7k*	*9k*	*11k*	*13k*

Find $P(X < 4)$, $P(X \geq 5)$, $P(3 < X \leq 6)$

[VTU July 2013, 2010]

Solution : Here, $p(x)$ is a valid pdf if (a) $P(x) \geq 0$ and

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

Hence we must have $k \geq 0$ and

$$k + 3k + 5k + 7k + 9k + 11k + 13k = 1$$

$$i.e, 49K = 1 \Rightarrow K = \frac{1}{49}$$

$$p(x < 4) = p(X = 0) + p(X = 1) + p(X = 2) + p(X = 3) = 16k = \frac{16}{49}$$

$$P(x \geq 5) = P(5) + P(6) = 11K + 13K = 24K = \frac{24}{49}$$

$$P(3 < x \leq 6) = P(4) + P(5) + P(6) = 33K = \frac{33}{49}$$

Problem 2. *A discrete random variable* X *has the p.m.f.:*

x	*1*	*2*	*3*	*4*	*5*	*6*	*7*
$f(x)$	k	$2k$	$2k$	$3k$	k^2	$2k^2$	$7k^2 + k$

(a) Find k *(b) Evaluate* $P(X < 3), P(X \geqslant 6)$

Solution: We know that $f(x)$ must satisfy two condition to qualify as the p.m.f.:

(i) $f(x) \geqslant 0 \forall x$

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

Now, we use the condition (ii) to evaluate k :

$\sum_{-\infty}^{\infty} f(x) = f(1) + f(2) + f(3) + f(4) + f(5) + f(6) + f(7) = 1$

$\Rightarrow k + 2k + 2k + 3k + k^2 + 2k^2 + 7k^2 + k = 1$

$\Rightarrow 10k^2 + 9k = 1$

$\Rightarrow 10k^2 + 10k - k - 1 = 0$

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

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

$\Rightarrow k = -1, \frac{1}{10}$

Now if we take $k = -1$, then condition (i) is not satisfied.

Taking $k = 1/10$ condition (i) becomes satisfied for all x. Next,

$$P(X < 3) = P\{(X = 1) + (X = 2)\} = P(X = 1) + P(X = 2)$$
$$= f(1) + f(2)$$
$$= \left(\frac{1}{10}\right) + 2\left(\frac{1}{10}\right) \quad \text{Remember that } f(1) = k, f(2) = 2k$$
$$= \frac{3}{10}$$
$$P(X \geq 6) = P(X = 6) + P(X = 7) = f(6) + f(7)$$
$$= 2\left(\frac{1}{100}\right)^2 + \left\{7\left(\frac{1}{10}\right)^2 + \left(\frac{1}{10}\right)\right\}$$
$$= 2\left(\frac{1}{100}\right) + \left\{7\left(\frac{1}{100}\right) + \left(\frac{1}{10}\right)\right\}$$
$$= \frac{2}{100} + \frac{7}{100} + \frac{10}{100} = \frac{19}{100}$$

Problem 3. *A random variable x has the following density function*

$$P(x) = \begin{cases} kx^2, & -3 \leq x \leq 3 \\ 0 & \text{elsewhere} \end{cases}$$

Evaluate k *and find (i)* $P(1 \leq x \leq 2)$ *(ii)* $P(x \leq 2)$ *(iii)* $P(x > 1)$

Solution : since p(x) is a pdf we must have

$P(x) \geq 0$ and $\int_{-\infty}^{\infty} P(x)dx = 1$

Now $P(x) \geq 0 \Rightarrow k \geq 0$ and

$$\int_{-\infty}^{\infty} P(x)dx = 1 \Rightarrow \int_{-3}^{3} kx^2 dx = 1$$

$$\Rightarrow \left[\frac{kx^3}{3}\right]_{-3}^{3} = 1 \Rightarrow k = \frac{1}{18}$$

$$P(1 \le x \le 2) = \int_1^2 \frac{x^2}{18}dx$$

$$= \frac{1}{18}\left[\frac{x^3}{3}\right]_1^2 = \frac{1}{54}(8-1) = \frac{7}{54}$$

$$P(x \le 2) = \int_{-3}^{2} \frac{1}{18}x^2 dx$$

$$= \frac{1}{18}\left[\frac{x^3}{3}\right]_{-3}^{2} = \frac{1}{54}(8+27) = \frac{35}{54}$$

$$P(x > 1) = \int_1^3 \frac{1}{18}x^2 dx$$

$$= \frac{1}{18}\left[\frac{x^3}{3}\right]_1^3 = \frac{1}{54}(27-1) = \frac{26}{54} = \frac{13}{27}$$

Problem 4. *The probability density function of a variate* x *is,*

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

Find k, *Mean, Variance,* SD *[VTU July 2023, Model 2020, 2004]*

Solution : Since $p(x)$ is a pdf, We must have (a) $P(x_i) \ge 0$ and

(b) $\sum P(x_i) = 1$

From (a) we have $k \ge 0$

From (b), we can write,

$$0.1 + k + 0.2 + 2k + 0.3 + k = 1 \Rightarrow 4K + 0.6 = 1 \Rightarrow K = 0.1$$

Mean is given by

$$(\mu) = \sum x_i P(x_i)$$
$$= (-2)(0.1) + (-1)k + (0)(0.2) + (1)(2k) + 2(0.3) + 3k$$
$$= -0.2 - 0.1 + 0.2 + 0.6 + 0.3 = 0.8$$

Variance is given by

$$(\sigma^2) = E[X^2] - \mu^2$$
$$= \sum x_i^2 P(x_i) - \mu^2$$
$$= (-2)^2(0.1) + (-1)^2 k + (0)^2(0.2) + (1)^2(2k) + 2^2(0.3) + 3^2 k - (0.8)^2$$
$$= (0.4 + 0.1 + 0.2 + 1.2 + 0.9) - (0.8)^2 = 2.16$$

Thus Mean $= 0.8$, Variance $= 2.16$ and S.D=$\sqrt{Var(X)} = \sqrt{2.16}$

Problem 5. *(i) Is the function defined by* $f(x) = e^{-x}, x > 0,\ f(x) = 0, x < 0$ *is a density function?*

(ii) If so, determine the probability that the variate having this density will fall in the interval $(1, 2)$

Solution : Here X is a continuous random variable.

(i) If $f(x)$ is density function, it must satisfy the properties,

(a) $f(x) \geq 0$, for all x.

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

Clearly We observe that condition (a) is satisfied ($\because f(x) \geq 0$.)

Now let us evaluate $\int_{-\infty}^{\infty} f(x)dx$

$$\int_{-\infty}^{\infty} f(x)dx = \int_{-\infty}^{0} f(x)dx + \int_{0}^{\infty} f(x)dx$$
$$= 0 + \int_{0}^{\infty} e^{-x}dx$$
$$= \left[-e^{-x}\right]_0^{\infty}$$
$$= -(0 - 1) = 1$$

i.e. condition (b) is satisfied.

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

(ii) The probability that the variate having this density will fall in the interval (1,2) is given by

$$\begin{aligned} P(1 < x < 2) &= \int_1^2 f(x)dx \\ &= -\left[e^{-x}\right]_1^2 \\ &= -\left(e^{-2} - e^{-1}\right) \\ &= 0.2325 \end{aligned}$$

Thus $P(1 < x < 2) = 0.2325$

Problem 6. *Find the constant* k *such that the function* $f(x) = \begin{cases} kx^2 & 0 \le x \le 3 \\ 0 & elsewhere \end{cases}$ *is a p.d.f. Also find (i)* $P(1 < X < 2)$ *(ii)* $P(X \le 1)$ *(iii)* $P(X > 1)$ *(iv) Mean (v) Variance* *[VTU July 2023, Jan 2018]*

Solution : Here, $f(x)$ is a valid pdf if

(a) $f(x) \ge 0$ and

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

From condition (a) we have $k \ge 0$

Using condition (b),

$$\int_{-\infty}^{\infty} f(x)dx = 1 \Rightarrow \int_0^3 kx^2dx = 1$$

$$\Rightarrow \left[\frac{kx^3}{3}\right]_0^3 = 1 \Rightarrow 9k = 1 \Rightarrow k = \frac{1}{9}$$

$$(i)\ P(1 < x < 2) = \int_1^2 f(x)dx$$
$$= \int_1^2 \frac{x^2}{9}dx$$
$$= \left[\frac{x^3}{27}\right]_1^2 = \frac{7}{27}$$

$$(ii)\ P(x \leq 1) = \int_0^1 f(x)dx$$
$$= \int_0^1 \frac{x^2}{9}dx$$
$$= \left[\frac{x^3}{27}\right]_0^1 = \frac{1}{27}$$

(iii)

$$P(x > 1) = \int_1^3 f(x)dx$$
$$= \int_1^3 \frac{x^2}{9}dx$$
$$= \left[\frac{x^3}{27}\right]_1^3 = \frac{26}{27}$$

(iv)

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

Mean, $= \int_0^3 x \cdot \frac{x^2}{9}dx$

$$= \left[\frac{x^4}{36}\right]_0^3 = \frac{81}{36} = \frac{9}{4}$$

(v) Variance

$$V = E[X^2] - \mu^2$$
$$= \left[\int_{-\infty}^{\infty} x^2 f(x)dx\right] - (\mu)^2$$
$$= \left[\int_0^3 x^2 \cdot \frac{x^2}{9}dx\right] - \left(\frac{9}{4}\right)^2$$
$$= \left[\frac{x^5}{45}\right]_0^3 - \frac{81}{16}$$
$$= \frac{81}{15} - \frac{81}{16} = \frac{81}{240} = \frac{27}{80}$$

Problem 7. *The pdf of the random variable* X *is given by the following table.*

x	-3	-2	-1	0	1	2	3
$P(x)$	K	$2k$	$3k$	$4k$	$3k$	$2k$	k

Find (i) k (ii) $P(X \leq 1)$ *(iii)* $P(X > 1)$ *(iv)* $P(-1 < X \leq 2)$ *(v) Mean of* X *(vi) SD of* X *[VTU: Dec/ Jan 16, July 2013]*

Solution : Since $p(x)$ is a pdf , we must have (a) $p(x) \geq 0$ for all x and (b) $\sum p(x) = 1$.

(i) From condition (a) we have $k \geq 0$

and the second Using condition (b), we have

$$k + 2k + 3k + 4k + 3k + 2k + k = 1 \text{ or } 16k = 1 \therefore k = \frac{1}{16}$$

(ii)

$$p(x \leq 1) = p(-3) + p(-2) + p(-1) + p(0) + p(1)$$
$$= 13k = \frac{13}{16}$$

(iii)

$$p(x > 1) = p(2) + p(3) = 3k = \frac{3}{16}$$

(iv)

$$p(-1 < x \leq 2) = p(0) + p(1) + p(2) = 9k = \frac{9}{16}$$

(v)

$$\text{Mean } = \mu = \sum x \cdot p(x)$$
$$= -3(k) - 2(2k) - 1(3k) + 0(4k) + 1(3k) + 2(2k) + 3(k)$$
$$= 0$$

(vi)

$$\text{Variance } = V = \sum (x - \mu)^2 \cdot p(x) = E(X^2) - \mu^2$$
$$= \sum x^2 \cdot p(x) - 0$$
$$= (-3)^2(k) + (-2)^2(2k) + (-1)^2(3k) + 0^2(4k)$$
$$+ 1^2(3k) + 2^2(2k) + 3^2(k)$$
$$= k(9 + 8 + 3 + 0 + 3 + 8 + 9) = \frac{1}{16}(40) = \frac{5}{2}$$

Thus $k = \frac{1}{16}$, Mean =0 and S . D .=$\sqrt{\frac{5}{2}}$

Problem 8. *The probability density $p(x)$ of a continuous random variable is given by $p(x) = y_0\, e^{-|x|}$, $-\infty < x < \infty$. Prove that $y_0 = \frac{1}{2}$. Find the mean and variance of the distribution.* *[vtu Dec 2012, 2004]*

Solution : Given pdf is $p(x) = y_0\, e^{-|x|}$, $-\infty < x < \infty$

Here, X is a continuous random variable.

Hence we must have,

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

$$\Rightarrow \int_{-\infty}^{\infty} y_0\, e^{-|x|}dx = 1$$

$$\Rightarrow 2\int_{0}^{\infty} y_0\, e^{-|x|}dx = 1 \qquad (\because e^{-|x|} \text{ is an even function})$$

$$\Rightarrow 2\int_{0}^{\infty} y_0\, e^{-x}dx = 1 \qquad (\because |x| = x, \text{ for } x > 0)$$

$$\Rightarrow 2y_0 \left[\frac{e^{-x}}{(-1)}\right]_0^{\infty} = 1$$

$$\Rightarrow 2y_0\,[0 - (-1)] = 1$$

$$\Rightarrow y_0 = \frac{1}{2}$$

Mean is,

$$\mu = \int_{-\infty}^{\infty} xp(x)dx$$

$$= \int_{-\infty}^{\infty} y_0\, xe^{-|x|}dx$$

$$= 0 \qquad (\because xe^{-|x|} \text{ is an odd function})$$

Now Variance is given by

$$\begin{aligned}
\sigma^2 &= E[X^2] - \mu^2 \\
&= \left[\int_{-\infty}^{\infty} x^2 p(x)dx\right] - \mu^2 \\
&= \left[\int_{-\infty}^{\infty} y_0\, x^2 e^{-|x|}dx\right] - 0 \qquad (\because x^2e^{-|x|} \text{ is an even function}) \\
&= y_0\, 2\int_0^{\infty} x^2 e^{-x}dx \\
&= y_0\, 2\left[(x^2)\frac{e^{-x}}{(-1)} - (2x)\frac{e^{-x}}{(-1)^2} + (2)\frac{e^{-x}}{(-1)^3}\right]_0^{\infty} \\
&= 2y_0[(0-0+0) - (0-0+-2)] \\
&= 2
\end{aligned}$$

Problem 9. *A random variable* X *has* $p(x) = 2^{-x}, x = 1, 2, 3\cdots$ *Show that* $p(x)$ *is a probability function. Also find* $p(\mathrm{X}$ *even* $)$, $p(\mathrm{X}$ *being divisible by* 3 *) and* $p(X \geq 5)$

Solution : Given pdf is $p(x) = 2^{-x} = \frac{1}{2^x}$

Clearly (a) $p(x) > 0$ forall x and

$$\begin{aligned}
(b)\sum_i p\,(x_i) &= \sum_{x=1}^{\infty}\frac{1}{2^x} = \frac{1}{2} + \frac{1}{2^2} + \frac{1}{2^3} + \cdots \\
&= \frac{1}{2}\left[1 + \frac{1}{2} + \frac{1}{2^2} + \cdots\right] \\
&= \frac{1}{2}\left(\frac{1}{1-\frac{1}{2}}\right) = 1
\end{aligned}$$

$(\because$ geometric series $1 + r + r^2 + \cdots = \dfrac{1}{1-r})$

Hence $p(x) = 2^{-x}$ is a probability function.

(i)

$$p(X\text{is even }) = p(X=2) + p(X=4) + p(X=6)\cdots$$
$$= \frac{1}{2^2} + \frac{1}{2^4} + \frac{1}{2^6} + \cdots$$
$$= \frac{1}{2^2}\left[1 + \frac{1}{2^2} + \left(\frac{1}{2^2}\right)^2 + \cdots\right]$$
$$= \frac{1}{2^2}\left[\frac{1}{1-\frac{1}{2^2}}\right]$$
$$= \frac{1}{4}\left(\frac{4}{3}\right) = \frac{1}{3}$$

Thus $p(X\text{is even }) = \frac{1}{3}$

(ii)

$$p(X\text{is divisible by } 3) = p(X=3) + p(X=6) + p(X=9) + \cdots$$
$$= \frac{1}{2^3} + \frac{1}{2^6} + \frac{1}{2^9} + \cdots$$
$$= \frac{1}{2^3}\left[1 + \frac{1}{2^3} + \left(\frac{1}{2^3}\right)^2 + \cdots\right]$$
$$= \frac{1}{2^3}\left[\frac{1}{1-\frac{1}{2^3}}\right]$$
$$= \frac{1}{2^3} \times \frac{8}{7} = \frac{1}{7}$$

Thus $p(X \text{is divisible by } 3) = \frac{1}{7}$

(iii)

$$p(X \geq 5) = 1 - p(X < 5)$$
$$= 1 - \sum_{i=1}^{4} p(x_i)$$
$$= 1 - \sum_{x=1}^{4} \frac{1}{2^x}$$
$$= 1 - \left(\frac{1}{2} + \frac{1}{2^2} + \frac{1}{2^3} + \frac{1}{2^4}\right)$$
$$= 1 - \left(\frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \frac{1}{16}\right)$$
$$= 1 - \frac{15}{16} = \frac{1}{16}$$

Thus $p(X \geq 5) = \frac{1}{16}$

Problem 10. *X is a continuous random variable with probability density function given by* $f(x) = \begin{cases} kx\,, & (0 \leq x < 2) \\ 2k\,, & (2 \leq x < 4) \\ -kx + 6k\,, & (4 \leq x < 6) \end{cases}$ *Find k and mean value of X*

Solution : Since $f(x)$ is a pdf, we must have

$$\int_0^6 f(x)dx = 1$$
$$\Rightarrow \int_0^2 kxdx + \int_2^4 2kdx + \int_4^6 (-kx + 6k)dx = 1$$
$$\Rightarrow k\left[x^2/2\right]_0^2 + 2k[x]_2^4 + \left(-kx^2/2 + 6kx\right)_4^6 = 1$$
$$\Rightarrow 2k + 4k + (-10k + 12k) = 1 \Rightarrow k = \frac{1}{8}$$

Mean of X is

$$\mu = \int_{-\infty}^{\infty} x\, f(x)dx$$
$$= \int_{0}^{6} xf(x)dx$$
$$= \int_{0}^{2} kx^2dx + \int_{2}^{4} 2kxdx + \int_{4}^{6} x(-kx+6k)dx$$
$$= k\left[\frac{x^3}{3}\right]_0^2 + 2k\left[\frac{x^2}{2}\right]_2^4 + \left(-k\left[\frac{x^3}{3}\right]_4^6 + 6k\left[\frac{x^2}{2}\right]_4^6\right)$$
$$= k\left(\frac{8}{3}\right) + k(12) - k\left(\frac{152}{3}\right) + 3k(20) = \frac{1}{8}(24) = 3$$

1.8 The Cumulative Distribution Function

The **cumulative distribution function (cdf)** of a random variable X denoted as $F(x)$, is the probability that a random variable X takes a value less than or equal to x.

Mathematically, it is defined as:

$F(x) = P(X \leq x)$

I

The CDF also has some important properties:

1. $F(x)$ is a non-decreasing function.
2. $0 \leq F(x) \leq 1$ for all x.

In the case of discrete random variable X,

$$F(x) = \sum_{-\infty}^{x} p(x)$$

and in the case of continuous random variable X,

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

n other words, the CDF is the integral of the PDF up to a certain point.

Probability in an Interval:

The probability that X falls in the interval (a, b) is given by the difference in the CDF values at b and a:

$$P(a < X < b) = F(b) - F(a) = \int_a^b f(x)\,dx$$

This is essentially the area under the PDF curve between a and b.

Problem 11. *Suppose that the error in the reaction temperature, in °C, for a controlled laboratory experiment is a continuous random variable X having the probability density function*

$$f(x) = \begin{cases} \frac{x^2}{3}, & -1 < x < 2 \\ 0, & \text{elsewhere} \end{cases}$$

(a) Verify that $f(x)$ is a density function.

(b) Find $P(0 < X \leq 1)$.

(c) Find the Cumulative density function.

Solution :(a) To verify $f(x)$ is a density function, we verify the following two conditions.

(i) non-negativity i.e., $f(x) \geq 0$, for all x.

(ii) Normality : i.e. $\int_{-\infty}^{\infty} f(x)dx = 1$

Obviously, $f(x) \geq 0$. and

$$\int_{-\infty}^{\infty} f(x)dx = \int_{-1}^{2} \frac{x^2}{3}dx$$
$$= \frac{x^3}{9}\Big|_{-1}^{2}$$
$$= \frac{8}{9} + \frac{1}{9} = 1$$

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

(b)
$$P(0 < X \leq 1) = \int_0^1 \frac{x^2}{3}dx$$
$$= \frac{x^3}{9}\Big|_0^1$$
$$= \frac{1}{9}.$$

(c) The Cumulative density function is

$$\begin{aligned} F(x) &= \int_{-\infty}^{x} f(t)dt \\ &= \int_{-1}^{x} \frac{t^2}{3}dt \\ &= \frac{t^3}{9}\Big|_{-1}^{x} \\ &= \frac{x^3+1}{9}. \end{aligned}$$

Therefore,

$$F(x) = \begin{cases} 0, & x < -1 \\ \frac{x^3+1}{9}, & -1 \le x < 2 \\ 1, & x \ge 2 \end{cases}$$

Problem 12. *A random variable X takes the values* $-3, -2, -1, 0, 1, 2, 3$ *such that* $P(X = 0) = P(X < 0)$ *and* $P(X = -3)$=$P(X = -2)$=$P(X = -1)$ = $P(X = 1)$ = $P(X = 2)$ = $P(X = 3)$. *Find the probability distribution. [VTU July 2017]*

Solution : Here X is a discrete random variable.

Let us denote probabilities of $X = -3, -2, -1, 0, 1, 2, 3$ respectively by $p_1, p_2, p_3, p_4, p_5, p_6, p_7$.

This can be represented as

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

By Data,

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

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

$$\Rightarrow P_4 = P_3 + P_2 + P_1 \qquad \cdots (1)$$

Also we have by data, $p_1 = p_2 = p_3 = p_5 = p_6 = p_7 = k$ (say)

Then (1) becomes $p_4 = k + k + k = 3k$

From the second property of pmf, we have

$$\sum_x p(x) = 1$$

$$p_1 + p_2 + p_3 + p_4 + p_5 + p_6 + p_7 = 1$$

$$k + k + k + 3k + k + k + k = 1$$

$$9k = 1$$

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

Substituting,

$$p_1 = p_2 = p_3 = p_5 = p_6 = p_7 = \frac{1}{9}$$

and

$$p_4 = 3k = \frac{3}{9}$$

Thus the probability distribution is as follows.

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

Problem 13. *If the random variable* X *takes the values* $1, 2, 3$ *and 4 such that* $2P(X = 1) = 3P(X = 2) = P(X = 3) = 5P(X = 4)$. *find the probability distribution function and cumulative distribution function of* X.

Solution : Let $P(X = 3) = k$, then

$$2P(X = 1) = k \quad \Rightarrow \quad P(X = 1) = \frac{k}{2}$$

$$3P(X = 2) = k \quad \Rightarrow \quad P(X = 2) = \frac{k}{3}$$

$$5P(X = 4) = k \quad \Rightarrow \quad P(X = 4) = \frac{k}{5}$$

$$\sum_{i=1}^{4} p\left(x_i\right) = 1 \Rightarrow \frac{k}{2} + \frac{k}{3} + k + \frac{k}{5} = 1$$

$$\frac{61k}{30} = 1 \Rightarrow k = \frac{30}{61}$$

The probability distribution function is

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

The CDF $F(x) = P(X \leq x)$ is defined as follows:

$If x < 1, F(x) = 0,$

If $1 \leq x < 2$, then $F(x) = P(X = 1) = \frac{15}{61}$,

If $2 \leq x < 3$, then $F(x) = P(X = 1) + P(X = 2) = \frac{25}{61}$,

If $3 \leq x < 4$, then

$$F(x) = P(X = 1) + P(X = 2) + P(X = 3) = \frac{55}{61},$$

If$x \geq 4$, then

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

Problem 14. *A coin is tossed twice. A random variable X represent the number of heads turning up. Find the discrete probability distribution for* X. *Also find its mean and variance.*

Solution : When a fair coin is tossed twice, the possible outcomes for each toss are either heads (H) or tails (T). The random variable X represents the number of heads turning up in the two tosses.

Here, Sample Space, $S = \{HH, HT, TH, TT\}$.

X= number of heads

The association of the elements of S to the random variable X are respectively $2, 1, 1, 0$

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

When $X = 0$, both tosses result in tails (TT). Hence $P(X = 0) = P(TT) = \frac{1}{4}$

When $X = 1$, One toss results in heads and the other in tails (HT or TH).

$P(X = 1) = P(HT \cup TH) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$

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

The discrete probability distribution for X is as follows.

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

Clearly, $p(x_i) > 0$ and $\sum p(x_i) = 1$

Mean $= \mu = \sum x_i\, p(x_i) = (0)\frac{1}{4} + (1)\frac{1}{2} + (2)\frac{1}{4} = 1$

$$E(X^2) = 0^2 \cdot \frac{1}{4} + 1^2 \cdot \frac{1}{2} + 2^2 \cdot \frac{1}{4} = \frac{3}{2}$$

Variance is

$$\text{Var}(X) = E(X^2) - [E(X)]^2 = \frac{3}{2} - 1 = \frac{1}{2}$$

Thus we have, Mean $= 1$ and Variance $= \frac{1}{2}$

Problem 15. *Find the constant k such that the function* $f(x) = \begin{cases} kxe^{-x} & 0 \le x \le 1 \\ 0 & elsewhere \end{cases}$ *is a p.d.f. Find the mean.* *(VTU Model 2022)*

Solution : Here, $f(x)$ is a valid pdf if

(i) $f(x) \ge 0$ and

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

From condition (i) we have $k \ge 0$

Using condition (ii),

$$\int_{-\infty}^{\infty} f(x)dx = 1 \Rightarrow \int_0^1 kxe^{-x}dx = 1$$

$$\Rightarrow k\int_0^1 xe^{-x}dx = 1$$

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

(using Bernoulli's rule for integration by parts)

$$\Rightarrow k\left[(-xe^{-x}) - e^{-x}\right]_0^1 = 1$$
$$\Rightarrow k\left[\{(-(1)e^{-1}) - e^{-1}\} - \{0 - e^{-0}\}\right] = 1$$
$$\Rightarrow k\left[(-2)e^{-1} + 1\right] = 1$$
$$\Rightarrow k = \frac{1}{1-2e^{-1}} = \frac{1}{1-\frac{2}{e}} = \frac{1}{\frac{e-2}{e}} = \frac{e}{e-2}$$

Mean is given by,

$$\mu = \int_{-\infty}^{\infty} x \cdot f(x)dx$$
$$= \int_0^1 x \cdot kxe^{-x}dx$$
$$= k\int_0^1 x^2e^{-x}dx$$
$$= k\left[x^2\frac{e^{-x}}{(-1)} - 2x\frac{e^{-x}}{(-1)^2} + 2\frac{e^{-x}}{(-1)^3}\right]_0^1$$
$$= k\left[(-1)e^{-1} - 2e^{-1} - 2e^{-1} - 0 - 0 + 2\right]$$
$$= k\left[(-5)e^{-1} + 2\right]$$
$$= k\left[\frac{-5}{e} + 2\right]$$
$$= k\left[\frac{2e-5}{e}\right]$$
$$= \frac{e}{e-2}\left[\frac{2e-5}{e}\right]$$
$$= \left[\frac{2e-5}{e-2}\right]$$

Problem 16. *Let X be the random variable that denotes the life in hours of a certain electronic device. The probability density function is*

$$f(x) = \begin{cases} \frac{20{,}000}{x^3}, & x > 100, \\ 0, & \text{elsewhere.} \end{cases}$$

Find the expected life of this type of device.

Solution : Here, X is a continuous random variable. Hence

$$\begin{aligned}\mu = E(X) &= \int_{-\infty}^{\infty} x\, f(x)dx \\ &= \int_{100}^{\infty} x\frac{20,000}{x^3}dx \\ &= \int_{100}^{\infty} \frac{20,000}{x^2}dx \\ &= \left[\frac{-20,000}{x}\right]_{100}^{\infty} \\ &= \left[0 - \frac{-20,000}{100}\right] = 200.\end{aligned}$$

Therefore, we can expect this type of device to last, on average, 200 hours.

Problem 17. *Let the random variable X represent the number of defective parts for a machine when 3 parts are sampled from a production line and tested. The following is the probability distribution of X.*

x	0	1	2	3
$f(x)$	0.51	0.38	0.10	0.01

calculate σ^2.

Solution : Here, X is a discrete randon variable.

First, we compute

$$\begin{aligned}\mu = E(X) &= \sum xf(x) \\ &= (0)(0.51) + (1)(0.38) + (2)(0.10) + (3)(0.01) \\ &= 0.61\end{aligned}$$

Now,

$$\begin{aligned}E\left(X^2\right) &= \sum xf(x) \\ &= (0)(0.51) + (1)(0.38) + (4)(0.10) + (9)(0.01) \\ &= 0.87.\end{aligned}$$

Therefore,

$$\begin{aligned}\sigma^2 &= E\left(X^2\right) - \mu^2 \\ &= 0.87 - (0.61)^2 \\ &= 0.4979.\end{aligned}$$

Problem 18. *The weekly demand for a drinking-water product, in thousands of liters, from a local chain of efficiency stores is a continuous random variable* X *having the probability density*

$$f(x) = \begin{cases} 2(x-1), & 1 < x < 2 \\ 0, & \text{elsewhere} \end{cases}$$

Find the mean and variance of X.

Solution : Here, X is a continuous random variable.

Calculating $E(X)$ and $E\left(X^2\right.$, we have

$$\begin{aligned} \mu = E(X) &= \int_{-\infty}^{\infty} x\, f(x)dx \\ &= 2\int_{1}^{2} x(x-1)dx \\ &= 2\int_{1}^{2} (x^2 - x)dx \\ &= 2\left[\frac{x^3}{3} - \frac{x^2}{2}\right]_1^2 \\ &= 2\left[\frac{2^3}{3} - \frac{2^2}{2} - \frac{1^3}{3} + \frac{1^2}{2}\right] \\ &= 2\left[\frac{8}{3} - 2 - \frac{1}{3} + \frac{1}{2}\right] \\ &= \frac{5}{3} \end{aligned}$$

and

$$\begin{aligned} E\left(X^2\right) &= \int_{-\infty}^{\infty} x^2\, f(x)dx \\ &= 2\int_{1}^{2} (x^3 - x^2)dx \\ &= 2\left[\frac{x^4}{4} - \frac{x^3}{3}\right]_1^2 \\ &= 2\left[\frac{2^4}{4} - \frac{2^3}{3} - \frac{1^4}{4} + \frac{1^3}{3}\right] \\ &= 2\left[4 - \frac{8}{3} - \frac{1}{4} + \frac{1}{3}\right] \\ &= \frac{17}{6}. \end{aligned}$$

Therefore,

$$\begin{aligned}\sigma^2 &= E\left(X^2\right) - \mu^2 \\ &= \frac{17}{6} - \left(\frac{5}{3}\right)^2 \\ &= \frac{1}{18}.\end{aligned}$$

Problem 19. *A variate* X *has the probability distribution*

x	:	-3	6	9
$P(X=x)$	:	$1/6$	$1/2$	$1/3$

Find $E(X)$ *and* $E\left(X^2\right)$. *Hence evaluate* $E(2X+1)^2$.

Solution : Here X is a discrete random variable.

$$\begin{aligned}E(X) &= \sum xp(x) \\ &= -3 \times \frac{1}{6} + 6 \times \frac{1}{2} + 9 \times \frac{1}{3} \\ &= 11/2.\end{aligned}$$

$$\begin{aligned}E(X)^2 &= \sum x^2 p(x) \\ &= 9 \times \frac{1}{6} + 36 \times \frac{1}{2} + 81 \times \frac{1}{3} \\ &= 93/2\end{aligned}$$

$$\begin{aligned}\therefore \quad E(2X+1)^2 &= E\left(4X^2 + 4X + 1\right) \\ &= 4E\left(X^2\right) + 4E(X) + 1 \\ &= 4(93/2) + 4(11/2) + 1 \\ &= 209.\end{aligned}$$

Bernoulli trials

- A random experiment having only two outcomes, arbitrarily labelled as success (S) and failure (F) is called a Bernoulli trial.

  Example 1: Result of an examination: “pass” (P) or “fail” (F).

  Example 2: While Tossing a coin if you define success as the outcome H and

Failure will be the outcome F.

- If p is the probability of success then $q = 1 - p$ is the probability of failure in a single Bernoulli trial.
  Always
  $$p + q = 1$$
  Example : In tossing a fair die, if you define success as getting the outcome 6 then getting other outcomes will be a Failure.
  Here $p = P(\{6\}) = \frac{1}{6}$ and $q = P(\{1,2,3,4,5\}) = \frac{5}{6}$
  Clearly, $p + q = 1$

1.9 Binomial Distribution

Suppose that we repeat Bernoulli trials n times independently under the same conditions. An experiment involving such independent Bernoulli trials is called a binomial experiment.

Let X= no. of successes in n trials.

Clearly X can take the values $X = 0, 1, 2, \cdots, n$

The probability of x successes out of n Bernoulli trials is given by,

$$P(x) = {}^nC_x p^x q^{n-x} \text{ with } p + q = 1$$

This probability distribution can be given as

x	0	1	2	$\cdots$	n
$p(x)$	q^n	${}^nC_1 p^1 q^{n-1}$	${}^nC_2 p^2 q^{n-2}$	$\cdots$	p^n

It may be observed that the value of $p(x)$ for different values $x = 0, 1, 2, \cdots, n$ are the successive terms in the binomial expansion of $(q + p)^n$ and accordingly this distribution is called the Binomial Distribution or Bernoulli Distribution.

The pdf of Binomial Random Variable is given by

$$\boxed{P(X = x) \text{ or } p(x) = {}^nC_x \, p^x q^{n-x}}$$

where

$n =$ no. of trials

$x =$ no. of successes in n trials

$p =$ Probability of Success in one trial

$q = 1 - p$ is the Probability of Failure in one trial

When X is a binomial random variable we also write it as $X \sim B(n, p)$

1.10 Mean and standard deviation of the Binomial Distribution :

The pmf of Binomial distribution is given by $p(x) = {}^nC_x p^x q^{n-x}, \ x = 0, 1, \cdots, n$

$$\text{Then } \mu = \mathrm{E(X)} = \sum_{x=0}^{n} x \ {}^nC_x p^x q^{n-x}$$

$$= \sum_{x=0}^{n} x \frac{n!}{(n-x)!x!} p^x q^{n-x}$$

$$= \sum_{x=1}^{n} \frac{n!}{(x-1)!(n-x)!} p^x q^{n-x}$$

(Since the $x = 0$ term vanishes)

$$\mu = \sum_{x=1}^{n} \frac{n(n-1)!}{(x-1)!(n-x)!}(p) p^{x-1} q^{n-x}$$

$$= np \sum_{x=1}^{n} \frac{(n-1)!}{(x-1)!(n-x)!} p^{x-1} q^{n-x}$$

$$= np \sum_{x=1}^{n} \frac{(n-1)!}{(x-1)![(n-1)-(x-1)]!} p^{x-1} q^{n-x}$$

$$= np \sum_{x=1}^{n} {}^{n-1}C_{x-1} p^{x-1} q^{n-x}$$

$$= np \left[{}^{n-1}\mathrm{C}_0 q^{n-1} + {}^{n-1}\mathrm{C}_1 p q^{n-2} + {}^{n-1}\mathrm{C}_2 p^2 q^{n-3} + \ldots + {}^{n-1}\mathrm{C}_{n-1} p^{n-1} \right]$$

$$= np \left[(p+q)^{n-1} \right]$$

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

But we know that $p + q = 1$

So $\mathrm{E(X)} = np(1)^{n-1} = np$

$$\mathrm{E(X^2)} = \sum_{x=0}^{n} x^2\, {}^nC_x p^x q^{n-x}$$

$$= \sum_{x=0}^{n} [x(x-1) + x]\, {}^nC_x p^x q^{n-x}$$

$$= \sum_{x=0}^{n} [x(x-1)]\, {}^nC_x p^x q^{n-x} + \sum_{x=0}^{n} x\, {}^nC_x p^x q^{n-x}$$

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

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

$$= n(n-1)p^2 \sum_{x=2}^{n} \frac{(n-2)!}{(x-2)![(n-2)-(x-2)]!} p^{x-2} q^{n-x} + np$$

$$= n(n-1)p^2 \sum_{x=2}^{n} {}^{n-2}C_{x-2} p^{x-2} q^{n-x} + np$$

$$= n(n-1)p^2 \left[{}^{n-2}\mathrm{C}_0 q^{n-2} + {}^{n-2}\mathrm{C}_1 pq^{n-3} + \ldots + {}^{n-2}\mathrm{C}_{n-2} p^{n-2}\right] + np$$

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

Since $p + q = 1$, we have

$$\mathrm{E}\left(\mathrm{X^2}\right) = n(n-1)p^2 + np$$

Using this,

$$\mathrm{Var(X)} = \mathrm{E}\left(\mathrm{X^2}\right) - [\mathrm{E(X)}]^2$$

$$= n(n-1)p^2 + np - (np)^2$$

$$= n^2p^2 - np^2 + np - n^2p^2$$

$$= np(1-p) = npq$$

$$S.D. = \sqrt{npq}$$

Using this result in (1) along with $\mu = np$ we have

$$Variance(V) = n(n-1)p^2 + np - (np)^2$$
$$= n^2p^2 - np^2 + np - n^2p^2 = np(1-p) = npq$$
$$Hence variance(V) = npq$$
$$S.D(\sigma) = \sqrt{V} = \sqrt{npq}$$

Thus we have for the binomial distribution,

Mean, $\mu = np$

$V(X)$ or $\sigma^2 = npq$ is the Variance,

$S.D.$ or $\sigma = \sqrt{npq}$ is the Standard Deviation

Problem 20. *Find n and p in the binomial distribution whose mean is 3 and variance is 2.*

Solution : Given, mean $= 3$ and variance is 2.

i.e $np = 3$ (1)

$npq = 2$ (2)

Dividing (2) by (1),

$$\frac{npq}{np} = \frac{2}{3}$$
$$\Rightarrow q = \frac{2}{3}$$
$$\therefore p = 1 - q = 1 - \frac{2}{3} = \frac{1}{3}$$
$$np = 3 \Rightarrow n\frac{1}{3} = 3 \Rightarrow n = 9$$

Problem 21. *A die is tossed thrice. A success is 'getting 1 or 6' on a toss. Find the mean and variance of the number of successes.* *[VTU Dec 2011]*

Solution : X = no. of successes

$p = \frac{2}{6} = \frac{1}{3}$

$q = 1 - p = \frac{2}{3}$

n= 3

Mean = $np = 3 \times \frac{1}{3} = 1$

Variance = $npq = (3)\frac{1}{3} \times \frac{2}{3} = \frac{2}{3}$

Problem 22. *The probability that a pen manufactured by a factory be defective is 1/10. If 12 such pens are manufactured, what is the probability that (i) exactly 2 are defective (ii) at least 2 are defective (iii) none of them are defective. [VTU :Jan 2020, July 2017, Dec/ Jan 16, , Dec 2012, 2004]*

Solution : Let X = no. of defective pens

$n = 12$ (no. of pens)

Probability of a defective pen is $p = 1/10 = 0.1$

Probability of a non-defective pen $= q = 1 - p = 1 - 0.1 = 0.9$

We have $P(x) = {}_{n}C_{x}\, p^x q^{n-x}$ where we have $n = 12$

(i) Prob. (exactly two defectives) is

$$P(x = 2) = {}_{12}C_{2}(0.1)^2(0.9)^{10} = 0.2301$$

(ii) Prob. (atleast 2 defectives) is,

$$\begin{aligned} P(X \geq 2) &= 1 - P(X < 2) \\ &= 1 - [P(x = 0) + P(x = 1)] \\ &= 1 - \left[{}^{12}C_0(0.1)^0(0.9)^{12} + {}^{12}C_1(0.1)^1(0.9)^{11}\right] \\ &= 0.341 \end{aligned}$$

(iii) Prob. (no defective) is ,

$$P(x = 0) = {}^{12}C_0(0.1)^0(0.9)^{12} = (0.9)^{12} = 0.2824$$

Problem 23. *The probability that a bomb dropped hits the target is 0.2, find the probability that out of 6 bombs dropped (i) Exactly two will hit the target (ii) Atleast 2 will hit the target [VTU:June/ July 15]*

Solution : Here let Success be hitting the target.

Hence X= no. of successes = no. of hits

$n = 6$ (no. of bombs),

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

and $\mathrm{q} = 1 - \frac{1}{5} = \frac{4}{5}$

We have $P(x) = {}_{n}C_{x}\, p^x q^{n-x}$

(i)P(Exactly two will hit the target) is

$$p(x=2) = n_{C_2} p^2 q^{n-2}$$
$$= {}^6C_2 \left(\frac{1}{5}\right)^2 \left(\frac{4}{5}\right)^4$$
$$= 0.2458$$

(ii) P (at least two bombs strike the target) is,

$$P(X \geq 2) = 1 - P(X < 2)$$
$$= 1 - [p(0) + p(1)]$$
$$= 1 - \left[n_{C_0} p^0 q^{n-0} + n_{C_1} p^1 q^{n-1}\right]$$
$$= 1 - \left[{}^6C_0 \left(\frac{1}{5}\right)^0 \left(\frac{4}{5}\right)^6 + {}^6C_1 \left(\frac{1}{5}\right)^1 \left(\frac{4}{5}\right)^5\right]$$
$$= 1 - 0.65536$$
$$= 0.3446$$

Problem 24. *In sampling a large no. of parts manufactured by a machine, the mean number of defectives in a sample of 20 is 2. Out of 1000 such samples, how many would be expected to contain atleast 3 defective parts.? [VTU June 2019, 2004]*

Solution :

In one sample :

let X= no. of defective parts.

n= no. of items = 20

Given that Mean, $(\mu) = np = 2$

$$\Rightarrow p = \frac{2}{n} = \frac{2}{20} = 0.1$$

Hence $q = 1 - p = 0.9$

The pdf is $P(x) = n_{C_x} p^x q^{n-x} = 20_{C_x}(0.1)^x(0.9)^{20-x}$ Probability of atleast 3 defective parts

$$P(X \geq 3) = P(3) + P(4) + \cdots + P(20)$$

OR

$$\begin{aligned} P(X \geq 3) &= 1 - P(X < 3) \\ &= 1 - [P(0) + P(1) + P(2)] \\ &= 1 - \left[(0.9)^{20} + 20_{C_1}(0.1)(0.9)^{19} + 20_{C_2}(0.1)^2(0.9)^{18}\right] \\ &= 0.323 \end{aligned}$$

Thus **in 1000 samples** the number of defectives is $1000 \times 0.323 = 323$

Problem 25. *Out of 800 families of 5 children each, how many would you expect to have (a) 3 Boys (b) 5 Girls (c) either 2 or 3 boys (d) atmost 2 girls? Assume equal probabilities for Boys and Girls.* *[VTU August 2022, August 2021, 2004]*

Solution : In one Family :

Let X= No. of Boys in the family

n= no. of children = 5 (given)

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

q = probability of having a girl $= \frac{1}{2}$

pdf of X is, $P(x) = n_{C_x} p^x q^{n-x} = 5_{C_x} \left(\frac{1}{2}\right)^x \left(\frac{1}{2}\right)^{5-x} = 5_{C_x} \frac{1}{2^5} = \frac{5_{C_x}}{32}$

In 800 Families : We have to find the no. of families with given conditions. Hence

(i) No. of Families with 3 Boys is,

$$\begin{aligned} 800P(X = 3) &= 800 \times \frac{5_{C_3}}{32} \\ &= 250 \end{aligned}$$

i.e. Expected number of families with 3 boys is 250.

(ii) No. of Families with 5 Girls = No. of Families with 0 Boys is given by

$$\begin{aligned} 800P(X = 0) &= 800 \times \frac{5_{C_0}}{32} \\ &= 25 \end{aligned}$$

i.e. Expected number of families with 5 girls is 25. (iii)No. of Families with 2 or 3 boys = No. of Families with 2 boys + No. of Families with 3 boys and is given by

$$800P(X=2)+800P(X=3)=800\times\frac{5_{C_2}}{32}+800\times\frac{5_{C_3}}{32}$$
$$=500$$

Expected number of faimilies with 2 or 3 boys is 500.

(iv) No. of Families with atmost 2 girls means that, families can have 5 boys and 0 girls or 4 boys and 1 girl or 3 boys and 2 girls.

Hence required answer is

$$800P(X=5)+800P(X=4)+800P(X=3)$$
$$=800\times\frac{5_{C_5}}{32}+800\times\frac{5_{C_4}}{32}+800\times\frac{5_{C_3}}{32}$$
$$=400$$

i.e. Expected number of families with atmost 2 girls is 400.

Problem 26. *The probability of germination of a seed in a packet of seeds is found to be 0.7. If 10 seeds are taken for experimenting on germination in a laboratory, find the probability that (i) 8 seeds germinate (ii) at least 8 seeds germinate (iii) at most 8 seeds germinate.* *[VTU Model 2020]*

Solution : Here, let X= no. of germinating seeds

n= Total no. of seeds = 10

p= P(probability of germination of a seed) = 0.7

$$q=1-p=0.3$$

The pdf of X is,

$$p(x)=n_{C_x}p^xq^{n-x}=10_{C_x}(0.7)^x(0.3)^{10-x}$$

(i) P(8 seeds germinate) is ,

$$P(X=8)=10_{C_8}(0.7)^8(0.3)^{10-8}$$
$$=0.2335$$

(ii) P(at least 8 seeds germinate) is,

$$\begin{aligned} P(X \geq 8) &= P(X=8) + P(X=9) + P(X=10) \\ &= 10_{C_8}(0.7)^8(0.3)^{10-8} + 10_{C_9}(0.7)^9(0.3)^{10-9} \\ &\quad + 10_{C_{10}}(0.7)^{10}(0.3)^{10-10} \\ &= 0.2335 + 0.1211 + 0.0282 \\ &= 0.3828 \end{aligned}$$

(iii) P(at most 8 seeds germinate) is

$$\begin{aligned} P(X \leq 8) &= 1 - P(X > 8) \\ &= 1 - [P(X=9) + P(X=10)] \\ &= 1 - [0.1211 + 0.0282] \\ &= 0.8507 \end{aligned}$$

1.11 Poisson Distribution

- Poisson distribution is regarded as the limiting form of the binomial distribution when n is very large $(n \to \infty)$ and p the probability of success is very small $(p \to 0)$
- np tends to a fixed finite constant say λ. (ie. $\lambda = np$)

The pdf of Poisson distribution is given by

$$\boxed{P(X=x) \text{ or } p(x) = \frac{e^{-\lambda}\lambda^x}{x!}}, \quad X = \{0, 1, 2, 3, \ldots\}$$

$P(x)$ is also called Poisson Probability function and x is called a Poisson Variate.

1.12 The Mean and Variance of Poisson Distribution

The pmf of Poisson distribution is

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

$$\text{Mean } \mathrm{E(X)} = \sum_{x=0}^{\infty} xp(x)$$
$$= \sum_{x=0}^{\infty} x\frac{e^{-\lambda}\lambda^x}{x!}$$
$$= \lambda e^{-\lambda}\left\{\sum_{x=1}^{\infty}\frac{\lambda^{x-1}}{(x-1)!}\right\}$$
$$= \lambda e^{-\lambda}\left\{1+\frac{\lambda}{1!}+\frac{\lambda^2}{2!}+\frac{\lambda^3}{3!}+\ldots\right\}$$
$$= \lambda e^{-\lambda}\left(e^{\lambda}\right) = \lambda$$
$$\mathrm{E}\left(\mathrm{X}^2\right) = \sum_{x=0}^{\infty} x^2p(x)$$
$$= \sum_{x=0}^{\infty} x\frac{e^{-\lambda}\lambda^x}{x!}$$
$$= \sum_{x=0}^{\infty}[x(x-1)+x]\frac{e^{-\lambda}\lambda^x}{x!}$$
$$= \sum_{x=0}^{\infty} x(x-1)\frac{e^{-\lambda}\lambda^x}{x!} + \sum_{x=0}^{\infty} x\frac{e^{-\lambda}\lambda^x}{x!}$$
$$= \sum_{x=2}^{\infty}\frac{e^{-\lambda}\lambda^x}{(x-2)!} + \mathrm{E(X)}$$
$$= e^{-\lambda}\lambda^2\sum_{x=2}^{\infty}\frac{\lambda^{x-2}}{(x-2)!} + \lambda$$
$$= e^{-\lambda}\lambda^2\left[1+\frac{\lambda}{1!}+\frac{\lambda^2}{2!}+\ldots..\right] + \lambda$$
$$= e^{-\lambda}\lambda^2\left(e^{\lambda}\right) + \lambda = \lambda^2 + \lambda$$
$$\mathrm{Var(X)} = \text{ Variance } (\mathrm{X}) = \mathrm{E}\left(\mathrm{X}^2\right) - [\mathrm{E(X)}]^2$$
$$= \lambda^2 + \lambda - \lambda^2 = \lambda$$
$$S.D. = \sqrt{\lambda}$$

The Mean and Variance of Poisson Distribution are given by

$$\boxed{\text{Mean, } \mu = \lambda, \quad \text{Variance, } V(X) or\ \sigma^2 = \lambda}$$

Problem 27. *If the mean of the poisson distribution is* 2. *Find* $\mathrm{P(x=0)}$

Solution: We know $P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}$

Given $\lambda = 2$

$$P(X = 0) = \frac{e^{-2}2^0}{0!}$$
$$= e^{-2} = 0.1353$$

Problem 28. *In a poisson distribution if* $P(x = 3) = P(x = 2)$ *find* $P(x = 0)$

Solution: Given: $P(x = 3) = P(x = 2)$ We know $P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}$

$$P(x = 3) = P(x = 2) \quad \text{(given)}$$
$$\Rightarrow \frac{e^{-\lambda}\lambda^3}{3!} = \frac{e^{-\lambda}\lambda^2}{2!}$$
$$\Rightarrow \frac{\lambda}{6} = \frac{1}{2}$$
$$\Rightarrow \lambda = 3$$
$$\therefore P(x = 0) = \frac{e^{-3}3^0}{0!}$$
$$= e^{-3} = 0.0497$$

Problem 29. *If* X *is a poisson variable with* $P(X = 2) = \frac{2}{3}P(X = 1)$, *find* $P(X = 3)$ *and* $P(X = 0)$

Solution: We know $P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}$ Given that

$$P(X = 2) = \frac{2}{3}P(X = 1)$$
$$\frac{e^{-\lambda}\lambda^2}{2!} = \frac{2}{3}\frac{e^{-\lambda}\lambda^1}{1!}$$
$$\frac{\lambda}{2} = \frac{2}{3}$$
$$\lambda = \frac{4}{3}$$

$$\therefore P(X=3) = \frac{e^{-\frac{4}{3}}\left(\frac{4}{3}\right)^3}{3!}$$
$$= 0.1041$$
$$P(X=0) = \frac{e^{-\frac{4}{3}}(\frac{4}{3})^4}{0!}$$
$$= e^{-\frac{1}{3}}$$
$$= 0.2635$$

Problem 30. *Given that 2% of fuses manufactured by a firm are defective, find by using Poisson distribution, the probability that a box containing 200 fuses has (i) no defective fuses (ii) 3 or more defective fuses (iii) At least one defective fuse [VTU August 2022, Dec 2018, Dec 2010]*

Solution: Here p = probability of a defective fuse $= 2\% = \frac{2}{100} = 0.02$
n= no. of fuses = 200
mean number of defectives $\lambda = np = 200 \times 0.02 = 4$ The Poisson distribution is given by

$$P(x) = \frac{\lambda^x e^{-\lambda}}{x!} = \frac{4^x e^{-4}}{x!}$$

(i) Probability of no defective fuse is

$$= P(0) = \frac{4^0 e^{-4}}{0!} = 0.0183$$

(ii) Probability of 3 or more defective fuses,

$$P(X \geq 3) = 1 - P(X < 3)$$
$$= 1 - [P(0) + P(1) + P(2)]$$
$$= 1 - \left[\frac{4^0 e^{-4}}{0!} + \frac{4^1 e^{-4}}{1!} + \frac{4^2 e^{-4}}{2!}\right]$$
$$= 1 - e^{-4}\left[1 + \frac{4^1}{1!} + \frac{4^2}{2!}\right]$$
$$= 1 - 0.0183(1 + 4 + 8) = 0.76189$$

(iii) Probability of at least one defective fuse is :

$$\begin{aligned} P(X \geq 1) &= 1 - P(X < 1) \\ &= 1 - p(X = 0) \\ &= 1 - 0.0183 \\ &= 0.9817 \end{aligned}$$

Problem 31. *In a certain factory turning out razor blades there is a small probability of 1/500 for any blade to be defective. The blades are supplied in packets of 10. Use poisson distribution to calculate the approximate number of packets containing (i)no defective (ii) One defective (ii) Two defectives in a consignment of 10000 packets*

[VTU: Model 2023, Model 2020, Dec/ Jan 16, 2004]

Solution : Here let X= no. of defective blades

p = probability of a defective blade $= \frac{1}{500} = 0.002$

In one packet :

no. of blades, $n = 10$ and

The mean number of defective blades is

$\lambda = np = 10 \times 0.002 = 0.02$

Poisson distribution is $P(x) = \frac{\lambda^x e^{-\lambda}}{x!} = \frac{e^{-0.02}(0.02)^x}{x!}$

In a consignment of 10000 packets :

(i) no. of packets with no defective is

$$10000 \times p(X = 0) = 10000 \times \frac{e^{-0.02}(0.02)^0}{0!}$$
$$= 9802$$

(i) no. of packets with one defective is

$$10000 \times p(X = 1) = 10000 \times \frac{e^{-0.02}(0.02)^1}{1!}$$
$$= 196$$

(iii) no. of packets with two defectives is

$$10000 \times p(X = 2) = 10000 \times \frac{e^{-0.02}(0.02)^2}{2!}$$
$$\approx 2$$

Problem 32. *A car hire -firm has two cars which it hires out on a day to day basis. The number of demands for a car is known to be Poisson distribution with mean* $\mathbf{1.5}$ *Find the probability of day on which (i) There is no demand for the car and (ii) The demand is rejected.*

Solution: Let X no. of demands for a car

n= no. of cars=2

Given that, $\boldsymbol{\lambda = 1.5}$

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

(i) The probability of day on which there is no demand for the car is

$$P(0) = \frac{(1.5)^0 e^{-1.5}}{0!} = e^{-1.5} = 0.2231$$

(ii)The demand will be rejected if no. of demands is more than no. of available cars. The probability that the demand is rejected is,

$$\begin{aligned} P(X > 2) &= 1 - P(X \leq 2) \\ &= 1 - [P(0) + P(1) + P(2)] \\ &= 1 - \left[\frac{(1.5)^0 e^{-1.5}}{0!} + \frac{(1.5)^1 e^{-1.5}}{1!} + \frac{(1.5)^2 e^{-1.5}}{2!}\right] \\ &= 1 - \left[e^{-1.5} + 1.5e^{-1.5} + 1.125e^{-1.5}\right] = 0.1912 \end{aligned}$$

Problem 33. *The number of accidents in a year to taxi drivers in a city follows a Poisson distribution with mean* $\mathbf{3}$*. Out of 1000 taxi drivers find approximately the number of the drivers with" (i) no accident in a year (ii) more than 3 accidents in a year.* *(VTU July 2023)*

Solution : For one driver:

Let X no. of accidents in a year

Given that X follows a Poisson distribution.

By data, mean $\boldsymbol{\lambda = 3}$

The Poisson distribution is given by

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

Out of 1000 drivers :

(i) Number of drivers with no accident in a year is,

$$1000 \times p(0) = 1000 \times \frac{e^{-3}\, 3^0}{0!} = 49.78 \approx 50$$

(ii) Number of drivers with more than 3 accidents in a year

$$\begin{aligned} 1000 \times p(X > 3) &= 1000 \times \{1 - P(x \leq 3)\} \\ &= 1000\,\{1 - [P(0) + P(1) + P(2) + P(3)]\} \\ &= 1000\left\{1 - \left[e^{-3}\left(\frac{3^0}{0!} + \frac{3^1}{1!} + \frac{3^2}{2!} + \frac{3^3}{3!}\right)\right]\right\} \\ &= 1000\,\{1 - [0.0498(1 + 3 + 4.5 + 4.5)]\} \approx 350 \end{aligned}$$

Extra problems from Binomial and Poisson Distributions

Problem 34. *The probability that a man aged 60 will live up to 70 is 0.65. What is the probability that out of 10 men, now 60, at least 7 will live to be 70? (VTU July 2023)*

Solution: Let X = No. of men aged 60 will live to be 70

p = The probability that a man aged 60 will live to be 70 = 0.65

$q = 1 - P = 1 - 0.65 = 0.35$

Total Number of men, (n) = 10

Since n is small, we shall use Binomial Distribution.

$$p(x) = {}^nC_x\, p^x q^{n-x} = {}^{10}C_x\, (0.65)^x (0.35)^{10-x}$$

and $X = 0, 1, 2, \ldots 10$

Probability that at least 7 men will live to 70,

$$\begin{aligned} P(X \geq 7) &= p(7) + p(8) + p(9) + p(10) \\ &= {}^{10}C_7\,(0.65)^7(0.35)^{10-7} \\ &\quad + {}^{10}C_8\,(0.65)^8(0.35)^{10-8} \\ &\quad + {}^{10}C_9\,(0.65)^9(0.35)^{10-9} \\ &\quad + {}^{10}C_{10}\,(0.65)^{10}(0.35)^{10-10} \\ &= 0.51383 \end{aligned}$$

Problem 35. *If the probability of a bad reaction from certain injection is 0.001, determine the chance that out of 2000 individuals more than two will get a bad reaction.* *[VTU July 2023, Jan 2018, 2008]*

Solution: Given that p= probability of a bad reaction = 0.001

n= no. of individuals, $n = 2000$ As the probability p is very small, this follows Poisson distribution.

Here Mean, $\lambda = np = 2000 \times 0.001 = 2$

Poisson distribution is

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

P(more than two will get a bad reaction) is,

$$\begin{aligned} P(x > 2) &= 1 - P(x \leq 2) \\ &= 1 - [P(x = 0) + P(x = 1) + P(x = 2)] \\ &= 1 - \left[\frac{e^{-2}(2)^0}{0!} + \frac{e^{-2}(2)^1}{1!}\frac{e^{-2}(2)^2}{2!}\right] \\ &= 1 - e^{-2}[1 + 2 + 2] \\ &= 1 - 5e^{-2} = 0.3233 \end{aligned}$$

Problem 36. *Suppose 300 misprints are randomly distributed throughout a book of 500 pages, find the probability that a given page contains (i) exactly three misprints (ii) less than three misprints and (iii) four or more misprints.* *[VTU Model 2020]*

Solution: we assume that X= the number of misprints on one page is the number of successes.

n = Total no. of misprints = 300 and $p = 1/500$

(assuming that misprints are evenly distributed in 500 pages)

Since n is large and p is small, we shall use Poisson Distribution.

Average no. of misprints is, $\lambda = np = \frac{300}{500} = 0.6$

Hence the pdf is $p(x) = \frac{\lambda^x e^{-\lambda}}{x!} = \frac{(0.6)^x e^{-(0.6)}}{x!}$

(i)P(exactly three misprints) is

$$p(X = 3) = \frac{(0.6)^3 e^{-(0.6)}}{3!} = 0.0198$$

(ii) p(less than three misprints) is

$$\begin{aligned} p(X < 3) &= p(0) + p(1) + p(2) \\ &= \frac{(0.6)^0 e^{-(0.6)}}{0!} + \frac{(0.6)^1 e^{-(0.6)}}{1!} + \frac{(0.6)^2 e^{-(0.6)}}{2!} \\ &= e^{-(0.6)} \left[1 + (0.6) + \frac{(0.6)^2}{2!}\right] \\ &= 0.5488 \left[1 + (0.6) + \frac{(0.6)^2}{2!}\right] \\ &= 0.9768 \end{aligned}$$

and (iii) p(four or more misprints) is,

$$\begin{aligned} p(X \geq 4) &= 1 - p(X < 4) \\ &= 1 - [p(0) + p(1) + p(2) + p(3)] \\ &= 1 - \left[\frac{(0.6)^0 e^{-(0.6)}}{0!} + \frac{(0.6)^1 e^{-(0.6)}}{1!} \right. \\ &\qquad \left. + \frac{(0.6)^2 e^{-(0.6)}}{2!} + \frac{(0.6)^3 e^{-(0.6)}}{3!}\right] \\ &= 1 - 0.9966 \\ &= 0.0034 \end{aligned}$$

Problem 37. *A die is thrown 8 times. Find the probability that 3 falls (i) Exactly two times (ii) At least once (iii) At the most 7 times* *[VTU July 2013]*

Solution: Let success be getting the outcome as 3

Here no of trials, $n = 8$

In each throw chance of 3 falls is 1 out of $\mathbf{6}$.

Therefore $p = \frac{1}{6},\ \ q = \frac{5}{6}$

The pdf is, $P(x) = {}^8C_x \left(\frac{1}{6}\right)^x \left(\frac{5}{6}\right)^{(8-x)}$

(i) The probability that 3 falls exactly 2 times $= P(2) = {}^8C_2 \left(\frac{1}{6}\right)^2 \left(\frac{5}{6}\right)^{(8-2)} = 0.2605$

(ii) The probability that 3 falls at least once is

$= P(X \geq 1) = 1 - P(0) = 1 - {}^8C_0 \left(\frac{1}{6}\right)^0 \left(\frac{5}{6}\right)^{(8-0)} = 0.7674$

(iii) The probability that 3 falls at the most 7 times is

$= P(X \leq 7) = 1 - P(8) = 1 - {}^8C_8 \left(\frac{1}{6}\right)^8 \left(\frac{5}{6}\right)^{(8-8)} = 0.9999 \approx 1$

Problem 38. *If the mean and standard deviation of the number of correctly answered questions in a test given to 4096 students are 2.5 and $\sqrt{1.875}$. Find an estimate of the number of candidates answering correctly (i) 8 or more questions (ii) 2 or less (iii) 5 questions.*

Solution: For a poisson distribution, mean = Variance = λ

We have mean $= 2.5$ and Variance = square of the S.D. = 1.875

Clearly, we can't use Poisson Distribution.

For a binomial distribution, by data $np = 2.5$ and $\sqrt{npq} = \sqrt{1.875}$ or $npq = 1.875$

Hence we have $2.5q = 1.875 \quad \therefore \quad q = 0.75; p = 1 - q = 0.25$

Now, $np = 2.5 \Rightarrow (0.25)n = 2.5 \Rightarrow n = 10$

For one student :

n = no. of questions= 10

Let x denote the number of correctly answered questions.

$$P(x) = n_{C_x} p^x q^{n-x} = 10_{C_x} (0.25)^x (0.75)^{10-x}$$

Out of 4096 students :

(i) Number of students correctly answering 8 or more questions is,

$$\begin{aligned}4096\, p(X \geq 8) &= 4096[p(8) + p(9) + p(10)] \\ &= 4096\left[10_{C_8}(0.25)^8(0.75)^{10-8} + 10_{C_9}(0.25)^9(0.75)^{10-9}\right. \\ &\qquad \left. +10_{C_{10}}(0.25)^{10}(0.75)^{10-10}\right] \\ &= 1.703 \approx 2\end{aligned}$$

(ii) number of candidates answering 2 or less questions correctly is,

$$\begin{aligned}4096\, p(X \leq 2) &= 4096\,[(p(0) + p(1) + p(2)] \\ &= 4096\left[10_{C_0}(0.25)^0(0.75)^{10-0} + 10_{C_1}(0.25)^1(0.75)^{10-1}\right. \\ &\qquad \left. +10_{C_2}(0.25)^2(0.75)^{10-2}\right] \\ &= 2152.8 \approx 2153\end{aligned}$$

i.e. No. of students correctly answering 2 or less than 2 questions is 2153.

(iii) number of candidates answering 5 questions correctly is,

$$\begin{aligned}4096\, p(5) &= 4096\left[10_{C_5}(0.25)^5(0.75)^{10-5}\right] \\ &= 239.2 \approx 239\end{aligned}$$

i.e. Number of students correctly answering 5 questions is 239.

1.13 Exponential distribution

The continuous probability distribution having the probability density function $f(x)$ given by

$$f(x) = \begin{cases} \alpha e^{-\alpha x} & \text{for } x \geq 0 \\ 0 & \text{otherwise , where } \alpha > 0 \end{cases}$$

is known as the exponential distribution.

Mean and Variance of this distribution is given by

$$\text{Mean}(\mu) = \frac{1}{\alpha};\quad \text{Variance}(\sigma^2) = \frac{1}{\alpha^2}$$

Problem 1.13.1. *If x is an exponential variate with mean 3 find (i) $P(x > 1)$ (ii) $P(x < 3)$*

Solution: The p.d.f of the exponential distribution is given by

$$f(x) = \begin{cases} \alpha e^{-\alpha x}, & 0 < x < \infty \\ 0 & \text{otherwise} \end{cases}$$

The mean of this distribution is given by $\frac{1}{\alpha}$

By data, miean $= \frac{1}{\alpha} = 3 \therefore \alpha = \frac{1}{3}$

$$\text{Hence } f(x) = \begin{cases} \frac{1}{3}e^{-x/3}, 0 < x < \infty \\ 0, \quad \text{otherwise} \end{cases}$$

(i)

$$\begin{aligned} P(x > 1) &= 1 - P(x \leq 1) \\ &= 1 - \int_0^1 f(x)dx \\ &= 1 - \int_0^1 \frac{1}{3}e^{-\frac{x}{3}}dx \\ &= 1 + \left[e^{-\frac{x}{3}}\right]_0^1 = e^{-\frac{1}{3}} = 0.7165 \end{aligned}$$

(ii)

$$\begin{aligned} P(x < 3) &= \int_0^3 f(x)dx \\ &= \int_0^3 \frac{1}{3}e^{-\frac{x}{3}}dx \\ &= -\left[e^{-\frac{x}{3}}\right]_0^3 = 1 - \frac{1}{e} = 0.6321 \end{aligned}$$

Problem 1.13.2. *In a certain town, duration of a shower is exponentially distributed with mean 5 minutes. What is the probability that the shower will last for (i) 10 minutes or more (ii) less than 10 minutes (iii) Between 10 and 12 minutes [VTU June 2019, Jan 2014, July 2013]*

Solution: The p.d.f of the exponential distribution is given by $f(x) = \alpha e^{-\alpha x}, x > 0$ and the mean is $= \frac{1}{\alpha}$

By data, $\frac{1}{\alpha} = 5 \therefore \alpha = \frac{1}{5}$

Hence $f(x) = \frac{1}{5}e^{-\frac{x}{5}}$

(i)

$$P(x \geq 10) = \int_{10}^{\infty} \frac{1}{5} e^{-\frac{x}{5}} dx$$
$$= -\left[e^{-\frac{x}{5}}\right]_{10}^{\infty}$$
$$= -\left(0 - e^{-2}\right) = e^{-2} = 0.1353$$

(ii)

$$P(x < 10) = \int_{0}^{10} \frac{1}{5} e^{-\frac{x}{5}} dx$$
$$= -\left[e^{-\frac{x}{5}}\right]_{0}^{10}$$
$$= -\left(e^{-2} - 1\right) = 1 - e^{-2}$$
$$= 0.8647$$

(iii)

$$P(10 < x < 12) = \int_{10}^{12} \frac{1}{5} e^{-\frac{x}{5}} dx$$
$$= -\left[e^{-\frac{x}{5}}\right]_{10}^{12}$$
$$= -\left(e^{-\frac{12}{5}} - e^{-2}\right)$$
$$= 0.0446$$

Problem 1.13.3. *The length of telephone conversations in a booth follows exponential distribution. If the average telephone conversation is 5 minutes, what is the probability that a a random call made from the booth (i) ends less than 5 minutes (ii)between 5 and 10 minutes?* *[VTU Model 2020, Jan 2020, Jan 2018]*

Solution: We have $f(x) = \alpha e^{-\alpha x}, x > 0$; Mean $= \frac{1}{\alpha}$

By data, $\frac{1}{\alpha} = 5 \Rightarrow \alpha = \frac{1}{5}$

Hence $f(x) = \frac{1}{5} e^{-\frac{x}{5}}$ is the p.d.f.

(i)

$$P(x < 5) = \int_{0}^{5} f(x) dx = \int_{0}^{5} \frac{1}{5} e^{-\frac{x}{5}} dx$$
$$= -\left[e^{-\frac{x}{5}}\right]_{0}^{5} = 1 - \frac{1}{e} = 0.6321$$

$$P(5 < x < 10) = \int_5^{10} f(x)dx$$
$$= \int_5^{10} \frac{1}{5} e^{-\frac{x}{5}} dx$$
$$= -\left[e^{-x/5}\right]_5^{10}$$
$$= \frac{1}{e} - \left(\frac{1}{e^2}\right) = 0.2325$$

Problem 1.13.4. *The length of a telephone conversation has been exponentially distributed with mean of 2 minutes. Find the probability that a call (i) ends in more than 3 minutes (ii) ends in less than 4 minutes and (iii) takes between 3 and 5 minutes*

Solution: For Exponential distribution,

$$\mu = \frac{1}{\alpha} = 2, \Rightarrow \alpha = \frac{1}{2}$$

Probability density function is

$$f(x) = \begin{cases} \frac{1}{2}e^{-\frac{x}{2}}, & \text{for } x \geq 0 \\ 0, & \text{for } x < 0 \end{cases}$$

(i) The probability that the conversation may last for more than 3 minutes is,

$$P(X > 3) = \int_3^{\infty} f(x)dx$$
$$= - \left[e^{-\frac{x}{2}}\right]_3^{\infty} = e^{-\frac{3}{2}} = 0.2231$$

(ii) The probability that the conservation may last for less than 4 minutes is,

$$P(X < 4) = \int_{-\infty}^{4} f(x)dx$$
$$= \frac{1}{2}\int_0^4 e^{-\frac{x}{2}} dx$$
$$= -\left. e^{-\frac{x}{2}}\right|_0^4 = 1 - e^{-2} = 0.8647$$

(iii) The probability that the conservation may last between 3 and 5 minutes

$$P(3 < X < 5) = \int_3^5 f(x)dx$$
$$= -\left. e^{-\frac{x}{2}}\right|_3^5$$
$$= -e^{-\frac{5}{2}} + e^{-\frac{3}{2}} = 0.1410$$

Problem 1.13.5. *The daily turn over in a medical shop is exponentially distributed with* $\mathbf{Rs.6000}$ *as the average with a net profit of* 8%**.** *Find the probability that the net profit exceeds Rs. 500 on a randomly chosen day.*

Solution: Let x denote the random variable denoting the turn over per day.

$$\text{Given} \quad \frac{1}{\alpha} = 6000 \Rightarrow \alpha = \frac{1}{6000}$$

Let A be the turn over in rupees for which the net profit is Rs. 500**.** Given that net profit is 8% of the turn over.

Hence we have

$$\frac{8}{100} \times A = 500 \Rightarrow A = 6250$$

i.e To get net profit as Rs. 500 the required turn over 6250 rupees.

Since the profit exceeds Rs. 500 the turn over has to exceed Rs.6250.

Hence the probability that the net profit exceeds Rs. 500 is given by

$$\begin{aligned} P(x > 6250) &= 1 - P(x \leq 6250) \\ &= 1 - \int_0^{6250} p(x)dx \\ &= 1 - \int_0^{6250} \frac{1}{6000} e^{-\frac{1}{6000}x} dx \\ &= 0.353 \end{aligned}$$

Problem 1.13.6. *The sales per day in a shop is exponentially distributed with the average sale amounting. to Rs. 100 and net profit is* 8%*. Find the probability that the net profit exceeds Rs. 30 on two consecutive days.*

Solution : Let x be the random variable of the sale in the shop. since x is an exponential variate the p.d.f $f(x) = \alpha e^{-\alpha x}, x > 0$

$$\text{Mean} = 1/\alpha = 100 \quad \therefore \quad \alpha = 1/100 = 0.01$$

Hence $f(x) = 0.01e^{-0.01x}, x > 0$

Let A be the amount of sale for which profit is Rs 30. Given that profit rate is 8%.

$$\Rightarrow A \times \frac{8}{100} = 30 \therefore A = 375$$

Probability of profit exceeding Rs.30 is equal to

$$P(\text{Profit} > \textbf{Rs.30}) = \text{1- Prob (profit} \leq \textbf{Rs.30})$$
$$= 1 - \text{ Prob (sales} \leq \textbf{Rs.375})$$
$$= 1 - \int_0^{375} (0.01)e^{-0.01x}dx$$
$$= 1 + \left[e^{-0.01x}\right]_0^{375} = e^{-3.75}$$

Probability of profit exceeding Rs.30 on a single day is $e^{-3.75}$

For two consecutive days, Probability of profit exceeding Rs.30 is, $e^{-3.75} \times e^{-3.75} = 0.00055$

1.14 Normal (Gaussian) Distribution

The PDF of Normal distribution is

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < \infty$$

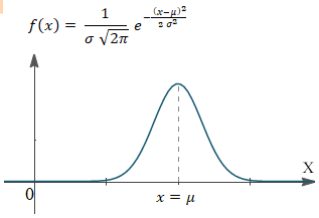
Mean $= \mu$

Variance $= \sigma^2$

Standard Deviation $= \sigma$

Normal curve

The normal curve is a bell-shaped curve symmetric about the line $x = \mu$ as shown in the following figure.

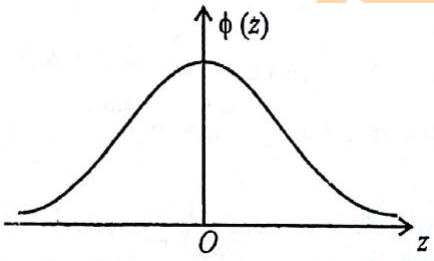


The line $x = \mu$ divides the total area under the curve which is equal to 1 into two equal parts. The area to the right as well as to the left of the line $x = \mu$ is 0.5

The pdf of Standard normal distribution is in the form :

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}, \quad -\infty < z < \infty$$

The standard normal curve is a bell-shaped curve symmetric about y-axis. Its mean is 0 and variance is 1.



For standardizing a Normal Random Variable, we use

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

The two tails of the standard normal probability distribution extend indefinitely and never touch the horizontal axis. The total area under this curve from $z = -\infty$ to ∞ is 1 unit.

By symmetry,

the area under the standard normal curve (from $z = -\infty$ to $z = 0$)= 0.5 and

Area from ($z = 0$ to $z = \infty$)=0.5

Evaluation of Probability in Normal Distribution :

In the case of normal distribution we have,

$$P(a \leq x \leq b) = \frac{1}{\sigma\sqrt{2\pi}} \int_a^b e^{-(x-\mu)^2/2\sigma^2} dx$$

This integral cannot be evaluated by known methods of integration and we have to employ the technique of numerical integration which becomes tedious.

Hence we think of standardization and the same.

To find any given probability of Normal random variable X, we convert X to Standard Normal variable z using $z = \frac{x-\mu}{\sigma}$

Hence $P(a \leq x \leq b)$ can be written as $P\left(\frac{a-\mu}{\sigma} \leq z \leq \frac{b-\mu}{\sigma}\right)$

If $\frac{a-\mu}{\sigma} = z_1$ and $\frac{b-\mu}{\sigma} = z_2$, then

$P(a < X < b)$

$= P(z_1 \leq z \leq z_2)$

$=$ Area from $z = z_1$ to $z = z_2$ under standard normal curve

$=$ Area from $(z = 0$ to $z = z_2) -$ Area from $(z = 0$ to $z = z_1)$

(under standard normal curve)

$= A(z_2) - A(z_1)$ (provided both z_1 and z_2 are > 0)

For $A(z)$ we use the Standard Normal table which gives Areas under the curve from 0 to z.

$$A(z) = \int_0^z \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz$$

represents the area under the standard normal curve from $z = 0$ to z

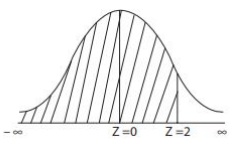
Key points of Normal Distribution

- $z = \frac{x-\mu}{\sigma}$
- $P(-\infty < x < \infty) = P(-\infty < z < \infty) = 1$
- $P(z \leq 0) = A(-\infty, 0) = A(0, \infty) = 0.5$
- $A(z)$ is the area of the curve area under the standard normal curve from $z = 0$ to z.

Problem 39. *If X is a normal variate with mean 80 and S.D 10. Compute* $P(X \leq 100)$.

Solution : We know $\mathbf{Z} = \frac{\mathbf{X}-\mu}{\sigma}$

When $X = 100$, $Z = \frac{100-80}{10} = 2$

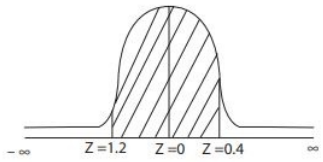


$$\therefore P(X \leq 100) = P(Z \leq 2)$$
$$= P(-\infty < Z < 0) + P(0 < Z < 2)$$
$$= 0.5 + 0.4772(\text{ from table })$$
$$= 0.9772$$

Problem 40. *If* X *is normally distributed with mean* 6 *and standard deviation* 5. *Find* $\mathrm{P}(0 \leq \mathrm{X} \leq 8)$

Solution :

$$\mathrm{P}(0 \leq \mathrm{X} \leq 8) = \mathrm{P}(-1.2 \leq \mathrm{Z} \leq 0.4)$$
$$= \mathrm{P}(-1.2 \leq \mathrm{Z} \leq 0) + \mathrm{P}(0 \leq \mathrm{Z} \leq 0.4)$$
$$= \mathrm{P}(0 \leq \mathrm{Z} \leq 1.2) + \mathrm{P}(0 \leq \mathrm{Z} \leq 0.4)$$
$$(by\ symmetry)$$
$$= 0.3849 + 0.1554 \quad \text{(from the table)}$$
$$= 0.5403$$



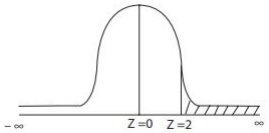
Problem 41. *When* X *is normally distributed with mean* 12, *standard deviation is* 4. *Find (i)* $\mathrm{P}(\mathrm{X} \geq 20)$ *(ii)* $P(0 < X < 12)$ *(iii)* $P(X \leq 20)$

Solution : Given: Mean $= \mu = 12$, S.D $= \sigma = 4$

We know $\mathrm{Z} = \frac{\mathrm{X}-\mu}{\sigma}$

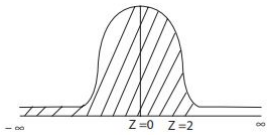
(i)

$$P(X \geq 20) = P(Z \geq 2)$$
$$= P(0 \leq Z \leq \infty) - P(0 \leq Z \leq 2)$$
$$= 0.5 - 0.4772 = 0.0228$$



(ii)

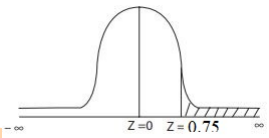
$$\mathrm{P}(0 < \mathrm{X} < 12) = \mathrm{P}(-3 < \mathrm{Z} < 0)$$
$$= P(0 < Z < 3) = 0.4987$$
$$\mathrm{P}(\mathrm{X} \leq 20) = \mathrm{P}(\mathrm{Z} \leq 2)$$
$$= \mathrm{P}(-\infty < \mathrm{Z} < 0) + \mathrm{P}(0 < \mathrm{Z} < 2)$$
$$= 0.5 + 0.4772 = 0.9772$$



Problem 42. *For the normal distribution with mean 2 and standard deviation 4, evaluate the following probabilities (i)* $P(X \geq 5)$ *(ii)* $P(|X| < 4)$ *(iii)* $P(|X| > 3)$

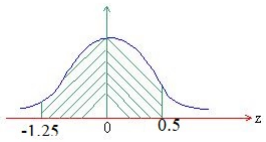
Solution : (i) $P(X \geq 5)$ when $X = 5, \quad z = \frac{5-2}{4} = \frac{3}{4} = 0.75$

$$P(X \geq 5) = P(z \geq 0.75)$$
$$= 0.5 - P(0 < z < 0.75)$$
$$= 0.5 - 0.2734$$
$$= 0.2266$$



(ii)

$$\begin{aligned}
P(|X| < 4) &= P(-4 < X < 4) \\
&= P\left(\frac{-4-\mu}{\sigma} < \frac{X-\mu}{\sigma} < \frac{4-\mu}{\sigma}\right) \\
&= P\left(\frac{-4-2}{4} < Z < \frac{4-2}{4}\right) \\
&= P(-1.5 < z < 0.5) = Area(-1.5 \text{ to } 0.5) \\
&= Area(-1.5 \text{ to } 0) + Area(0 \text{ to } 0.5) \\
&= Area(0 \text{ to } 1.5) + Area(0 \text{ to } 0.5) \\
&= 0.4332 + 0.1915 = 0.6247
\end{aligned}$$



(iii)

$$\begin{aligned}
P(|X| > 3) &= 1 - P(|X| \le 3) \\
&= 1 - P(-3 \le X \le 3) \\
&= 1 - P(-3 \le X \le 3) \\
&= 1 - P\left(\frac{-3-\mu}{\sigma} \le \frac{X-\mu}{\sigma} \le \frac{3-\mu}{\sigma}\right) \\
&= 1 - P\left(\frac{-3-2}{4} \le Z \le \frac{3-2}{4}\right) \\
&= 1 - P(-1.25 \le Z \le 0.25) \\
&= 1 - Area(-1.25 \text{ to } 0.25) \\
&= 1 - [Area(-1.25 \text{ to } 0) + Area(0 \text{ to } 0.25)] \\
&= 1 - [Area(0 \text{ to } 1.25) + Area(0 \text{ to } 0.25)] \\
&= 1 - [0.3944 + 0.0987] = 0.5069
\end{aligned}$$

Problem 43. *In a test on electric bulbs, it was found that the life time of a particular brand distributed normally with an average life of 2000 hours and S.D. of 60 hours. If a firm purchases 2500 bulbs find the number of bulbs that are likely to last for*

(i) more than 2100 hours (ii) less than 1950 hours (iii) between 1900 to 2100 hours [VTU: July 2023, Model 2020, June/July 17]

Solution : By data $\mu = 2000, \sigma = 60$

We have standard normal variable, $z = \frac{x-\mu}{\sigma} = \frac{x-2000}{60}$

(i) To find $P(x > 2100)$ If $x = 2100, z = 100/60 \approx 1.67$

$$\begin{aligned} P(x > 2100) &= P(z > 1.67) \\ &= P(z \geq 0) - P(0 < z < 1.67) \\ &= 0.5 - \phi(1.67) \\ &= 0.5 - 0.4525 = 0.0475 \end{aligned}$$

Hence number of bulbs that are likely to last for more than 2100 hours is

$$2500 \times 0.0475 = 118.75 = 119$$

(ii) To find $P(x < 1950)$

If $x = 1950, z = -5/6 = -0.83$

$$\begin{aligned} P(x < 1950) &= P(z < -0.83) \\ &= P(z > 0.83) \\ &= P(z \geq 0) - P(0 < z < 0.83) \\ &= 0.5 - \phi(0.83) \\ &= 0.5 - 0.2967 = 0.2033 \end{aligned}$$

Hence number of bulbs that are likely to last for less than 1950 hours is

$$2500 \times 0.2033 \approx 508$$

(iii) To find $P(1900 < x < 2100)$ If $x = 1900, z = -1.67$ and if $x = 2100, z = 1.67$

$$\begin{aligned} P(1900 < x < 2100) &= P(-1.67 < z < 1.67) \\ &= 2P(0 < z < 1.67) \\ &= 2\phi(1.67) = 2 \times 0.4525 = 0.905 \end{aligned}$$

Hence number number of bulbs that are likely to last between 1900 and 2100 hours

$$= 2500 \times 0.905 = 2262.5 \approx 2263$$

Problem 44. *The marks of 1000 students in an examination follows a normal distribution with mean 70 and standard deviation of 5. Find the expected no. of students, whose marks will be (i)less than 65 (ii) more than 75 (iii)Between 65 and 75. Given (A(z=1)= 0.3413) [VTU August 2021, Model 2020, Jan 2014, Dec 2010]*

Solution : Let x represent the marks of students. By data $\mu = 70, \sigma = 5$
Hence standard Normal variable is $z = \frac{x-\mu}{\sigma} = \frac{x-70}{5}$
(i) If $x = 65, z = -1$ and we have to find $P(z < -1)$

$$P(z < -1) = P(z > 1)$$
$$= P(z \geq 0) - P(0 < z < 1)$$
$$= 0.5 - A(1) = 0.5 - 0.3413 = 0.1587$$

Number of students scoring less than 65 marks $= 1000 \times 0.1587 = 158.7 \approx 159$
(ii) If $x = 75, z = 1$ and we have to find $P(z > 1)$

$$P(z > 1) = P(z \geq 0) - P(0 < z < 1)$$
$$= 0.5 - A(1) = 0.5 - 0.3413 = 0.1587$$

$\therefore$ the number of students scoring more than 75 marks $= 1000 \times 0.1587 = 158.7 \approx 159$
(iii) We have to find $P(-1 < z < 1)$

$$p(-1 < z < 1) = 2p(0 < z < 1)$$
$$= 2A(1) = 2(0,343) = 0.6836$$

$\therefore$ number of students scoring marks between 65 and 75 $= 1000 \times 0.6826 = 682.6 \approx 683$

Problem 45. *In a test on 2000 electric bulbs, it was found that the life of a particular make was normally distributed with an average life of 2040 hours and SD of 60 hours. Estimate the number of bulbs likely to burn for (i) More than 2150 hours (ii) Less than 1950 hours (iii) More than 1920 hours but less than 2160 hours. Given A(1.5) = 0.4332, A(1.83) = 0.4664, A(2) = 0.4772 [VTU: Dec 2018, Dec14/ Jan 15]*

Solution : (i) p(More than 2150 hours) is,

$$p(X > 2150) = p\left(\frac{X-\mu}{\sigma} > \frac{2150-\mu}{\sigma}\right)$$
$$= p(Z > 1.833)$$
$$= Area(0 \text{ to } \infty) - Area(0 \text{ to } 1.833)$$
$$= 0.5 - 0.4664 = 0.0336$$

Thus number of bulbs likely to burn for more than 2150 hours = $2000 \times 0.0336 \approx 67$ (ii)P(bulb likely to burn for Less than 1950 hours) is

$$p(X < 1950) = p\left(\frac{X-\mu}{\sigma} < \frac{1950-\mu}{\sigma}\right)$$
$$= p(Z < -1.5)$$
$$= Area(-\infty \text{ to } -1.5)$$
$$= Area(1.5 \text{ to } \infty)$$
$$= Area(0 \text{ to } \infty) - Area(0 \text{ to } 1.5)$$
$$= 0.5 - 0.4332 = 0.0668$$

Thus number of bulbs likely to burn for Less than 1950 hours = $2000 \times 0.0668 \approx 134$

(iii) P(bulb likely to burn for more than 1920 hours but less than 2160 hours) is,

$$p(1920 < X < 2160) = p\left(\frac{1920-\mu}{\sigma} < \frac{X-\mu}{\sigma} < \frac{2160-\mu}{\sigma}\right)$$
$$= p(-2 < z < 2)$$
$$= 2 \times Area(0 \text{ to } 2)$$
$$= 2 \times 0.4772 = 0.9544$$

Thus number of bulbs likely to burn for more than 1920 hours but less than 2160 hours = $2000 \times 0.9544 = 1909$

Problem 46. *In a normal distribution 31% of items are under 45 and 8% of the items are over 64. Find the mean and S.D. of the distribution. [VTU: July 2023, Dec/ Jan 2016, Dec 2012, 2009]*

Solution : Let μ and σ be the mean and S.D of the normal distribution By data $P(x < 45) = 0.31$ and $P(x > 64) = 0.08$

Now

$$P(x < 45) = 0.31 \Rightarrow P\left(z < \frac{45-\mu}{\sigma}\right) = 0.31$$

$$\Rightarrow Area\left(-\infty \text{ to } \frac{45-\mu}{\sigma}\right) = 0.31$$

$$\Rightarrow Area\left(\frac{\mu-45}{\sigma} \text{ to } \infty\right) = 0.31$$

$$\Rightarrow 0.5 - Area\left(0 \text{ to } \frac{\mu-45}{\sigma}\right) = 0.31$$

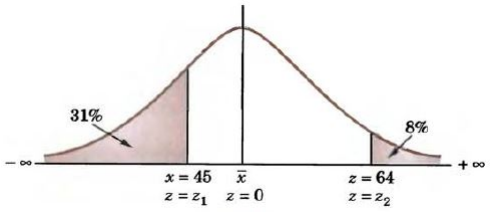
$$\Rightarrow A\left(\frac{\mu-45}{\sigma}\right) = 0.5 - 0.31 = 0.19$$

$$(But\ 0.19 \approx A(0.5))$$

$$\Rightarrow \left(\frac{\mu-45}{\sigma}\right) = 0.5$$

$$\Rightarrow \mu - 45 = 0.5\sigma$$

$$\Rightarrow \mu - 0.5\sigma = 45 \qquad (*)$$



$$P(x > 64) = 0.08 \Rightarrow P\left(z > \frac{64-\mu}{\sigma}\right) = 0.08$$

$$\Rightarrow Area\left(\frac{64-\mu}{\sigma} \text{ to } \infty\right) = 0.08$$

$$\Rightarrow Area(0 \text{ to } \infty) - Area\left(0 \text{ to } \frac{64-\mu}{\sigma}\right) = 0.08$$

$$\Rightarrow 0.5 - A\left(\frac{64-\mu}{\sigma}\right) = 0.08$$

$$\Rightarrow A\left(\frac{64-\mu}{\sigma}\right) = 0.5 - 0.08 = 0.42$$

$$(But\ 0.42 \approx A(1.4))$$

$$\Rightarrow \frac{64-\mu}{\sigma} = 1.4$$

$$\Rightarrow 64 - \mu = 1.4\sigma$$

$$\Rightarrow \mu + 1.4\sigma = 64 \qquad (**)$$

Now we have two equations with two unknowns. They are

$\mu - 0.5\sigma = 45 \qquad (*)$

and

$\mu + 1.4\sigma = 64 \qquad (**)$,

Solving these two, we get $\mu = 50, \sigma = 10$

Problem 47. *In an examination, 7% of students score less than* 35% *marks and 89% of students score less than 60% marks. Find the mean and standard deviation, if the marks are normally distributed. It is given that* $P(0 < z < 1.23) = 0.39$ *and* $P(0 < z < 1.48 = 0.43)$ *[Jan 2020, VTU Jan 2012]*

Solution : Let μ and σ be the mean and $S.D$ of the normal distribution.

By data we have $P(x < 35) = 0.07, P(x < 60) = 0.89$

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

When $x = 35, \quad z = \frac{35-\mu}{\sigma} = z_1$(say)

When $x = 60, \quad z = \frac{60-\mu}{\sigma} = z_2$(say)

Hence we have

$$P(x < 35) = 0.07 \Rightarrow P(z < z_1) = 0.07$$

$$\Rightarrow Area(-\infty \text{ to } z_1) = 0.07$$

$$\Rightarrow Area(-z_1 \text{ to } \infty) = 0.07$$

$$\Rightarrow 0.5 - A(-z_1) = 0.07$$

$$\Rightarrow A(-z_1) = 0.5 - 0.07 = 0.43$$

$$(But\ 0.43 \approx A(1.48))$$

$$\Rightarrow -z_1 = 1.48$$

$$\Rightarrow -\frac{35-\mu}{\sigma} = 1.48$$

$$\Rightarrow \mu - 1.48\sigma = 35 \qquad (*)$$

$$P(x < 60) = 0.89 \Rightarrow P(z < z_2) = 0.89$$

$$\Rightarrow Area(-\infty \text{ to } z_2) = 0.89$$

$$\Rightarrow Area(-\infty \text{ to } 0) + Area(0 \text{ to } z_2) = 0.89$$

(Here the area measured from $-\infty$ to z_2 is more than 0.5)

(Hence z_2 lies right to y axis)

$$\Rightarrow 0.5 + A(Z_2) = 0.89$$

$$\Rightarrow A(z_2) = 0.89 - 0.5 = 0.39$$

$$(But\ 0.39 \approx A(1.23))$$

$$\Rightarrow z_2 = 1.23$$

$$\Rightarrow \frac{60-\mu}{\sigma} = 1.23$$

$$\Rightarrow \mu + 1.23\sigma = 60 \qquad (**)$$

Solving,

$\mu - 1.48\sigma = 35 \qquad (*)$

and

$\mu + 1.23\sigma = 60 \qquad (**)$

we get $\mu = 48.65$ and $\sigma = 9.23$

1.15 Question Bank

Random Variables

1. The pdf of the random variable x is given by the following table.

x	-3	-2	-1	0	1	2	3
P(x)	K	2k	3k	4k	3k	2k	k

Find (i) k (ii) $P(X \leq 1)$ (iii) $P(X > 1)$ (iv) $P(-1 < X < 2)$ (v) Mean of X (vi) SD of X

[VTU: Dec/ Jan 16, July 2013]

Ans : $k = \frac{1}{16}, \frac{13}{16}, \frac{3}{16}, \frac{9}{16}, 0, \sqrt{\frac{5}{2}}$

2. A random variable x has the following distribution. Find k, SD and mean of the distribution.

x	-2	-1	0	1	2	3	4
P(x)	0.1	0.1	k	0.1	2k	k	k

Find the value of k and calculate mean and Standard Deviation. [VTU: June/ July 15, Dec 2012, 2004]

3. The probability density function of a variate x is,

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

Find k, Mean, Variance, SD [VTU July 2023, Model 2020, 2004] Ans : $0.1, 0.8, 2.16, 1.47$

4. A random variable X takes the values $-3, -2, -1, 0, 1, 2, 3$ such that $P(X = 0) = P(X < 0)$ and $P(X = -3) = P(X = -2) = P(X = -1)$=

$P(X = 1) = P(X = 2)$=$P(X = 3)$. Find the probability distribution.

[VTU July 2017]

Ans :

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

5. The probability density function of a random variable X is

X	0	1	2	3	4	5	6
p(X) :	k	3k	5k	7k	9k	11k	13k

(i) Find $P(X < 4)$, $P(X \geq 5)$, $P(3 < X \leq 6)$ [VTU July 2013, 2010]

Ans : $\frac{16}{49}, \frac{24}{49}, \frac{33}{49}$

6. The probability density function of a random variable X is

X	0	1	2	3	4	5
p(X) :	k	3k	5k	7k	9k	11k

(i) Find k (ii) $P(3 < x <\leq 5)$ [VTU Dec 2018]

7. A random variable X has the following probability density 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) Find the value of k (ii) Evaluate $P(X < 3)$, $P(X \geq 6)$ (iii) $P(3 < X \leq 6)$ [July 2023, Jan 2020, VTU Jan 2014] Ans : $\frac{3}{10}, \frac{19}{100},$

8. The probability density $p(x)$ of a continuous random variable is given by $p(x) = y_0 e^{-|x|}$, $-\infty < x < \infty$. Prove that $y_0 = \frac{1}{2}$. Find the mean and variance of the distribution. [vtu Dec 2012, 2004]

Ans : $\frac{1}{2}$, $\mu = 0, \sigma^2 = 2$

9. The probability density $p(x)$ of a continuous random variable is given by $p(x) = y_0 e^{-|x|}$, $-10 < x < \infty$. Find y_0. Also find the mean. [VTU Jan 2015]

10. The frequency function of a continuous random variable is given by $f(x) = y_0\ x(2 - x)$, $0 \leq x \leq 2$. Find the value of y_0, mean and variance of X.

11. If $f(x) = \begin{cases} \frac{1}{2}(x+1), & -1 < x < 1 \\ 0, & elsewhere \end{cases}$ represents the density function of a random variable X, Find $E(X)$ and $Var(X)$

12. Is the function defined by $f(x) = e^{-x}, x > 0, f(x) = 0, x < 0$ is a density function?
    (i) If so, determine the probability that the variate having this density will fall in the interval (1,2).
    (ii) Also, find the cumulative probability function $F(2)$ [VTU July 2017]

13. Find the constant k such that the function $f(x) = \begin{cases} kx^2 & 0 \le x \le 3 \\ 0 & elsewhere \end{cases}$ is a p.d.f. Also find (i) $P(1 < X < 2)$ (ii) $P(X \le 1)$ (iii) $P(X > 1)$ (iv) Mean (v) Variance [VTU Jan 2018] Ans : $k = \frac{1}{9}$, (i)$\frac{7}{27}$, (ii)$\frac{1}{27}$, (iii)$\frac{26}{27}$, (iv) $\frac{9}{4}$, (v) $\frac{27}{80}$

14. Find the constant 'c' such that the function $p(x) = \begin{cases} cx^2 & 0 \le x \le 3 \\ 0 & elsewhere \end{cases}$ is a p.d.f. Also find (i) $P(1 < X < 2)$ (ii) $P(X \le 1)$ (iii) $P(X > 1)$ [VTU Model 2020, Jan 2018]

15. Suppose a random variable X takes the values -3, -1, 2 and 5 with respective probabilities $\frac{2k-3}{10}, \frac{k-2}{10}, \frac{k-1}{10}, \frac{k+1}{10}$. Find the value of k and (i) $P(-3 < X < 4)$ (ii) $P(X \le 2)$

    [VTU June 2010]

16. The probability density function of a variate x is,

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

    Find k, Mean, Variance, SD [VTU Model 2020, 2004]

17. The probability distribution of a random variable X is given by the following table:

$X\ (=x_i)$	0	1	2	3	4	5
$P(X)$	k	$5k$	$10k$	$10k$	$5k$	k

Find (i) the value of k (ii) $P(x \leq 1)$ (iii) $P(0 \leq x < 3)$ [VTU Model 2020]

18. A random variable x has the following density function

$$P(x) = \begin{cases} kx^2, & -3 \leq x \leq 3 \\ 0 & \text{elsewhere} \end{cases}$$

Evaluate k and find (i) $P(1 \leq x \leq 2)$ (ii) $P(x \leq 2)$ (iii) $P(x > 1)$ [VTU Model 2020]

19. If the random variable X takes the values $1, 2, 3$ and 4 such that $2P(X=1) = 3P(X=2) = P(X=3) = 5P(X=4)$. find the probability distribution function and cumulative distribution function of X.

20. A coin is tossed twice. A random variable X represent the number of heads turning up. Find the discrete probability distribution for X. Also find its mean and variance. Ans : Mean $= 1$ and Variance $= \frac{1}{2}$

21. A random variable X has $p(x) = 2^{-x}, x = 1, 2, 3 \cdots$ Show that $p(x)$ is a probability function. Also find $p(\text{X even})$, $p(\text{X being divisible by } 3)$ and $p(X \geq 5)$ Ans : $\frac{1}{3}, \frac{1}{7}, = \frac{1}{16}$

22. The total number of hours, measured in units of 100 hours, that a family runs a vacuum cleaner over a period of one year is a continuous random variable X that has the density function

$$f(x) = \begin{cases} x, & 0 < x < 1, \\ 2 - x, & 1 \leq x < 2, \\ 0, & \text{elsewhere} \end{cases}$$

Find the probability that over a period of one year, a family runs their vacuum cleaner (a) less than 120 hours; (b) between 50 and 100 hours.

(c) Find the average number of hours per year that families run their vacuum cleaners. Ans : $(a)0.68; (b)0.375, (c)100$

23. Consider the density function

$$f(x) = \begin{cases} k\sqrt{x}, & 0 < x < 1, \\ 0, & \text{elsewhere} \end{cases}$$

(a) Evaluate k. (b) Find $F(x)$ and use it to evaluate

$$P(0.3 < X < 0.6).$$

Ans : $(a) 3/2; (b) F(x) = \begin{cases} 0\sqrt{x}, & x < 0, \\ x^{\frac{3}{2}}, 0 \leq x < 1 \\ 1, x \geq 1 \end{cases}$

24. The random variable X, representing the number of errors per 100 lines of software code, has the following probability distribution:

x	2	3	4	5	6
$f(x)$	0.01	0.25	0.4	0.3	0.04

find the variance of X. Ans : 0.74

25. A shipment of 8 similar microcomputers to a retail outlet contains 3 that are defective. If a school makes a random purchase of 2 of these computers, find the probability distribution for the number of defectives. Find the mean and variance of this distribution. (VTU Model 2024)

Binomial, Poisson Distributions

1. Derive the expressions for mean and variance of binomial distribution. [VTU July 2023, Dec 2018, Jan 2018, July 2017, Jan 2015, Dec 2012, June 2012]

2. The probability of germination of a seed in a packet of seeds is found to be 0.7. If 10 seeds are taken for experimenting on germination in a laboratory, find the probability that (i) 8 seeds germinate (ii) at least 8 seeds germinate (iii) at most 8 seeds germinate. [VTU Model 2020] Ans : 0.2335, 0.3828, 0.8507

3. Obtain the Mean and variance of the Poisson distribution. [VTU :July 2023, Jan 2020, June 2019, July 2017, Dec/ Jan 14]

4. The probability that a pen manufactured by a factory be defective is 1/10. If 12 such pens are manufactured, what is the probability that (i) exactly 2 are defective (ii) at least 2 are defective (iii) none of them are defective. [VTU :Jan 2020, July 2017, Dec/ Jan 16, , Dec 2012, 2004] Ans : 0.2301, 0.341, 0.2824

5. In a certain factory turning out razar blades there is a small probability of 1/500 for any blade to be defective. The blades are supplied in packets of 10. Use poisson distribution to calculate the approximate number of packets containing (i) One defective (ii) Two defective, in a consignment of 1000 packets

[VTU:Model 2020, Dec/ Jan 16, 2004]

Ans : 9802, 196, 2

6. The probability that a bomb dropped hits the target is 0.2, find the probability that out of 6 bombs dropped (i) Exactly two will hit the target (ii) Atleast two will hit the target [VTU:June/ July 15] Ans : 0.2458, 0.3446

7. A certain screw making machine produces on average two defective out of 100 and packs them in boxes of 500 . Find the probability that the box contains

i) three defective

ii) At least one defective

iii) between 2 and 4 defective. [VTU: Feb 2021]

Ans : $\mathbf{0.2277.0.0198, 0.057}$

8. A die is tossed thrice. A success is 'getting 1 or 6' on a toss. Find the mean and variance of the number of successes. [VTU Dec 2011] Ans : Mean = 1, Variance= 2/3

9. A die is thrown 8 times. Find the probability that 3 falls (i) Exactly two times (ii) At least once (iii) At the most 7 times [VTU July 2013] Ans : 0.2605, 0.7674, $\mathbf{0.9999 \approx 1}$

10. In sampling a large no. of parts manufactured by a machine, the mean number of defectives in a sample of 20 is 2. Out of 1000 such samples, how many would be

expected to contain atleast 3 defective parts.? [VTU June 2019, 2004] Ans : 323

11. Out of 800 families of 5 children each, how many would you expect to have (a) 3 Boys (b) 5 Girls (c) either 2 or 3 boys ? (d) atmost 2 girls ? Assume equal probabilities for Boys and Girls.

[VTU August 2022, August 2021, 2004]

Ans :a) 250 , b)25, c)500 d)400

12. In a bombing action, there are 50% chance that any bomb will strike the target. Two direct hits are required to destroy the target completely. How many bombs are required to be dropped to give a 99% chance or better of completely destroying the target? [VTU 2003]

13. If the probability of a bad reaction from certain injection is 0.001, determine the chance that out of 200 individuals more than two will get a bad reaction. [VTU July 2023, Jan 2018, 2008] Ans : 0.32

14. The probability that an individual suffers a bad reaction from a certain injection is 0.001. Using Poisson distribution, determine the probability that out of 2000 individuals: (i) Exactly 3 and (ii) More than 2 , will suffer a bad reaction. [VTU July 2023, June 2012]

15. If a random variable has Poisson Distribution such that $p(1) = p(2)$, find (i) Mean of the Distribution (ii) $P(4)$ [VTU 2003]

16. If 10% of rivets produced by a machine are defective. Find the probability that, out of 12 such rivets, (i)exactly 2 are defective (ii) at least 2 are defective (iii) none of them are defective. [VTU July 2013]

17. Given that 2% of fuses manufactured by a firm are defective, find by using Poisson distribution, the probability that a box containing 200 fuses has (i) no defective fuses (ii) 3 or more defective fuses (iii) At least one defective fuse [VTU August 2022,, Dec 2018, Dec 2010] Ans : 0.0183, 0.76189, 0.9817

18. Suppose 300 misprints are randomly distributed throughout a book of 500 pages, find the probability that a given page contains (i) exactly three misprints (ii) less than three misprints and (iii) four or more misprints. [VTU Model 2020] Ans : 0.0198, 0.9768, 0.0034

19. If $\mathbf{20\%}$ bolts produced by a machine are defective. Calculate the probability that out of 7 randomly selected bolts not more than 1 is defective, atmost two are defective? [VTU August 2021]

20. In a poisson distribution if $\boldsymbol{P(x=3)=P(x=2)}$ find $\boldsymbol{P(x=0)}$ Ans : 0.0497

21. If $\boldsymbol{X}$ is a poisson variable with $\boldsymbol{P(X=2)=\frac{2}{3}P(X=1)}$, find $\boldsymbol{P(X=3)}$ and $\boldsymbol{P(X=0)}$ Ans : 0.1041, 0.2635

22. A car hire -firm has two cars which it hires out on a day to day basis. The number of demands for a car is known to be Poisson distribution with mean $\mathbf{1.5}$ Find the probability of day on which (i) There is no demand for the car and (ii) The demand is rejected. Ans : 0.2231, 0.1912

23. The number of accidents in a year to taxi drivers in a city follows a Poisson distribution with mean $\mathbf{3}$. Out of 1000 taxi drivers find approximately the number of the drivers with“ (i) no accident in a year (ii) more than 3 accidents in a year. (VTU July 2023)

Ans : 50, 350

24. The probability that a man aged 60 will live up to 70 is 0.65. What is the probability that out of 10 men, now 60, at least 7 will live to be 70? (VTU July 2023)

Ans : 0.51383

25. If the mean and standard deviation of the number of correctly answered questions in a test given to 4096 students are 2.5 and $\sqrt{\mathbf{1.875}}$. Find an estimate

of the number of candidates answering correctly (i) 8 or more questions (ii) 2 or less (iii) 5 questions. Ans : $\mathbf{1.703 \approx 2, 2152.8 \approx 2153, 239.2 \approx 239}$

26. In a quiz contest of answering 'Yes' or 'No', what is the probability of guessing atleast 6 answers correctly out of 10 questions asked? Also find the probability of the same if there are 4 options for a correct answer? (VTU July 2023)

Ans : 0.377, 0.0197

27. The probability that a news reader commits no mistake in reading the news is $\frac{1}{e^3}$. Find the probability that on a particular news broadcast he commits (i) Only 2 mistakes (ii) more than 3 mistakes (iii) atmost 3 mistakes, assuming that mistakes follow Poisson distribution. (VTU 2023)

Ans : $\boldsymbol{\lambda = 3, 0.2240, 0.3528, 0.6472}$

28. The number of telephone lines busy at an instant of time is a binomial variate with probability 0.1 that a line is busy. If 10 lines are chosen at random, what is the probability that, (i) no line is busy (ii) all lines are busy (iii) at least one line is busy (iv) Atmost 2 lines are busy. (VTU 2023)

Ans : $\mathbf{0.3487, (0.1)^10, 0.6513, 0.9298}$

Exponential and Normal Distributions

1. In a certain town, duration of a shower is exponentially distributed with mean 5 minutes. What is the probability that the shower will last for (i) 10 minutes or more (ii) less than 10 minutes (iii) Between 10 and 12 minutes [VTU August 2021, Model 2020, June 2019, Jan 2014, July 2013]
2. If X is an exponential variate with mean 4, evaluate (i) $\boldsymbol{P(0 < X < 1)}$ (ii) $\boldsymbol{P(X > 2)}$ and (iii) $\boldsymbol{P(-\infty < X < 10}$ [VTU Dec 2012]
3. If $\boldsymbol{x}$ is an exponential variate with mean 3 find (i) $\boldsymbol{P(x > 1)}$ (ii) $\boldsymbol{P(x < 3)}$
4. The length of telephone conversations in a booth follows exponential distribution. If the average telephone conversation is 5 minutes, what is the

probability that a a random call made from the booth (i) ends less than 5 minutes (ii)between 5 and 10 minutes? [VTU Model 2020, Jan 2020, Jan 2018]

5. At a certain city bus stop three buses arrive per hour on an average. Assuming that the time between successive arrivals is exponentially distributed, find the probability that the time between the arrival of successive buses is (i) less than 10 minutes and (ii) at least 30 minutes Ans: $\mathbf{0.3935, 0.2231}$

6. The length of a telephone conversation has been exponentially distributed with mean of 2 minutes. Find the probability that a call (i) ends in more than 3 minutes (ii) ends in less than 4 minutes and (iii) takes between 3 and 5 minutes Ans: $\mathbf{0.6321, 0.179}$

7. The length of a telephone conversation has been exponentially distribution with mean of 3 minutes. Find the probability that a call (i) ends in more than 1 minutes and (ii) takes less than 3 minutes. Ans: $\mathbf{0.71655, 0.63212}$

8. The daily turn over in a medical shop is exponentially distributed with $\mathbf{Rs.6000}$ as the average with a net profit of $\mathbf{8\%}$. Find the probability that the net profit exceeds Rs. 500 on a randomly chosen day.

9. The sales per day in a shop is exponentially distributed with the average sale amounting. to Rs. 100 and net profit is 8%. Find the probability that the net profit exceeds Rs. 30 on two consecutive days.

10. X is a normal variate with mean 30 and S.D, 5, find the probabilities that (i) $\mathbf{26 \leq X \leq 40}$ (ii) $\boldsymbol{X \geq 45}$ and (iii) $\mathbf{|X - 30| > 5}$ [VTU August 2022, June 2019]

Ans : 0.7653, 0.0014, 0.3174

11. If X is a normal variate with mean 80 and S.D 10. Compute $\boldsymbol{P(X \leq 100)}$.

Ans : 0.9772

12. In a test on electric bulbs, it was found that the life time of a particular brand distributed normally with an average life of 2000 hours and S.D. of 60 hours if a firm purchases 2500 bulbs had the member of bulbs that are likely to last for (i) more than 2100 hours (ii) less than 1950 hours (iii) between 1900 to 2100 hours [VTU Model 2020, June/July 17]

13. In a normal distribution 31% of items are under 45 and 8% of the items are over 64. Find the mean and S.D. of the distribution. [VTU: Dec/ Jan 2016, Dec 2012, 2009] Ans : 50 and 10

14. In a test on 2000 electric bulbs, it was found that the life of a particular make was normally distributed with an average life of 2040 hours and SD of 60 hours. Estimate the number of bulbs likely to burn for (i) More than 2150 hours (ii) Less than 1950 hours (iii) More than 1920 hours but less than 2160 hours. Given A(1.5) = 0.4332, A(1.83) = 0.4664, A(2) = 0.4772 [VTU: Dec 2018, Dec14/ Jan 15]

Ans : $\mathbf{0.0336 * 2000 = 67}$, $\mathbf{0.0918 * 2000 = 184}$, $\mathbf{0.9544 * 2000 = 1909}$

15. The life of an electric bulb is normally distributed with mean life of 200 hours and S.D. of 60 hours. Out of 2500 bulbs, find the number of bulbs which are likely to last between 1900 and 2100 hours. Given that $\boldsymbol{P(0 < Z < 1.67) = 0.4525}$ [VTU Dec 2018, July 2017, Jan 2014]

16. The marks of 1000 students in an examination follows a normal distribution with mean 70 and standard deviation of 5. Find the expected no. of students, whose marks will be (i)less than 65 (ii) more than 75 (iii)Between 65 and 75. Given (A(z=1)= 0.3413) [VTU August 2021, Model 2020, Jan 2014, Dec 2010]

Ans : $\mathbf{159.159, 683}$

17. 200 students appeared in an examination, distribution of marks is assumed to be normal with mean = $\mu = 30$, S.D.=$\sigma = 6.25$. How many students are expected to get marks (i) between 20 and 40 (ii) less than 35 [VTU July 2013]

18. In an examination, 7% of students score less than 35% marks and 89% of students score less than 60% marks. Find the mean and standard deviation, if the marks are normally distributed. It is given that $P(0 < z < 1.2263) = 0.39$ and $P(0 < z < 1.4757 = 0.43)$ [Jan 2020, VTU Jan 2012]

19. The weekly wages of workers in a company are normally distributed with mean of Rs. 700/- and standard deviation of Rs. 50. Find the probability that the weekly wage of a randomly chosen worker is (i) Rs. 650 and Rs. 750 (ii) More than Rs 750 [VTU June 2012]

20. A sample of 100 dry battery cells tested to find the length of life produced by a company and following results are recorded : Mean life = 12 hours, Standard Deviation = 3 hours. Assuming data to be normally distributed, find the expected life of a dry cell : (i) more than 15 hours (ii) Between 10 and 1 hours [VTU Jan 2011]

21. Suppose that the student IQ scores form a normal distribution with mean 100 and standard deviation 20. Find the percentage of students whose (i) score is less than 80 (ii) score falls between 90 and 140 (iii) score is more than 120 [VTU June 2011]

22. The I.Q. of students in a certain college is assumed to be normally distributed with mean 100 and variance 25. If two students are selected at random, find the probability that (i) both of them have I.Q. between 102 and 110 (ii) at least one of them have I.Q . between 102 and 110 (iii) at most one of them have I.Q. between 102 and 110. [VTU Model 2020]

Module 2

Joint probability distribution & Markov Chain

Syllabus :

Joint probability distribution: Joint Probability distribution for two discrete random variables, expectation, covariance and correlation.
Markov Chain: Introduction to Stochastic Process, Probability Vectors, Stochastic matrices, Regular stochastic matrices, Markov chains, Higher transition probabilities, Stationary distribution of Regular Markov chains and absorbing states.

2.1 Joint probability and Joint probability distribution

- If X and Y are two discrete random variables, we define the joint probability function of X and Y by $P(X = x, Y = y) = f(x, y)$ (or $p(x, y)$), where $f(x, y)$ satisfy the conditions

  (i) $p(x, y) \geq 0, \quad \forall x \in X$ and $y \in Y$

  (ii) $\sum_{x,y} p(x, y) = 1$

- The joint Probabilities $P(x_i, y_j) = P(X = x_i, Y = y_j)$ are presented in a two way table called Joint Probability Table.

X\Y	y_1	y_2	...	y_n	Marginal Density of X
x_1	$p(x_1, y_1)$	$p(x_1, y_2)$	...	$p(x_1, y_n)$	$P(x_1)$
x_2	$p(x_2, y_1)$	$p(x_2, y_2)$	...	$p(x_2, y_n)$	$P(x_2)$
x_3	$p(x_3, y_1)$	$p(x_3, y_2)$	...	$p(x_3, y_n)$	$P(x_3)$
.	.	.		.	.
.	.	.		.	.
.	.	.		.	.
x_m	$p(x_m, y_1)$	$p(x_m, y_2)$	...	$p(x_m, y_n)$	$P(x_m)$
Marginal Density of Y	$P(y_1)$	$P(y_2)$	...	$P(y_n)$	Sum=1

From joint pmf of X and Y we can find the pmf (or pdf) of X without Y. This is called a marginal pmf (or pdf).

In the table each row represents a single value of X and each column represents a single value of Y.

- Marginal Densities of X (i.e. $p(x_i)$) are obtained by adding the corresponding entries of the x_i-row and we put the sum in the right-hand margin of the table.

$$p_X(x_i) = \sum_j p(x_i, y_j)$$

- Marginal Densities of Y(i.e. $p(y_j)$) are obtained by adding the corresponding entries of the - y_j column and we put the sum in the bottom margin of the table.

$$p_Y(y_j) = \sum_i p(x_i, y_j)$$

2.2 Independent Random Variables

Two random Variables are said to be stochastically independent if they satisfy the condition

$$p(x_i, y_j) = p(x_i)\, p(y_j)$$

for all $x_i \in X$ and $y_j \in Y$

2.3 Expectation, Variance, Covariance and Correlation

If X is a discrete random variable taking values $x_1, x_2, \ldots x_n$ having probability function $p(x)$ then the Expectation of X (denoted by $E(X)$ or μ_X) is defined by the relation

$$\mu_x = E(X) = \sum xp(x)$$

The Variance of X denoted by $V(X)$ is defined by the relation

$$V(X) \text{ or } \sigma_X^2 = \sum_{i=1}^{n} (x_i - \mu)^2 p(x_i) = E(X^2) - (E(X))^2$$

Where μ is the mean of x, $\sigma_x = \sqrt{V(X)}$ is called the standard deviation (S.D) of X.

If X and Y are random variables having the joint probability function $p(x, y)$, then the expectation of X and Y are defined as follows :

$$E(X) \text{ or } \mu_X = \sum xp(x),$$

$$E(Y) \text{ or } \mu_Y = \sum yp(y) \text{ and}$$

$$E(XY) = \sum x\, y\, p(x, y)$$

The covariance of X and Y denoted by $COV(X, Y)$ is defined by the relation,

$$\text{COV(X, Y)} = \sum_i \sum_j (x_i - \mu_x)(y_j - \mu_y) p(x_i, y_j)$$

$$= \text{E}[(\text{X} - \mu_X)(Y - \mu_Y)]$$

$$\text{COV}(X, Y) = E(XY) - \mu_X \mu_Y \text{ (Equivalently)}$$

Further, the coefficient of correlation of X and Y denoted by $\rho(\text{X}, \text{Y})$ is defined by the relation

$$\rho(X, Y) = \frac{\text{COV}(X, Y)}{\sigma_x \sigma_y}$$

Problem 48. *The joint distribution of two random variables X and Y is as follows.*

$X \backslash Y$	-4	2	7
1	$\frac{1}{8}$	$\frac{1}{4}$	$\frac{1}{8}$
5	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{8}$

Determine (i) Marginal Distributions of X and Y (ii) Covariance of X and Y (iii) σ_x and σ_y (iv)Correlation of X and Y *[VTU Dec 2018, July 2017]*

Solution : (i)The distributions (marginal distribution) of X and Y are obtained by adding the all the respective row entries and also the respective column entries.

$$P(X=1)=\frac{1}{8}+\frac{1}{4}+\frac{1}{8}=\frac{1}{2}$$
$$p(X=5)=\frac{1}{4}+\frac{1}{8}+\frac{1}{8}=\frac{1}{2}$$
$$p(Y=-4)=\frac{1}{8}+\frac{1}{4}=\frac{3}{8}$$
$$p(Y=2)=\frac{1}{4}+\frac{1}{8}=\frac{3}{8}$$
$$p(Y=7)=\frac{1}{8}+\frac{1}{8}=\frac{1}{4}$$

Hence distributions of X and Y are given by

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

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

(ii)

$$\mathrm{E(X)}=\sum \mathrm{x_i p\,(x_i)}=(1)(1/2)+(5)(1/2)=3$$
$$\mathrm{E(Y)}=\sum \mathrm{y_j p\,(y_j)}=(-4)(3/8)+(2)(3/8)+(7)(1/4)=1$$

Thus

$$\mu_X=\mathrm{E(X)}=3 \text{ and } \mu_y=E(Y)=1$$
$$\mathrm{E(XY)}=\Sigma \mathrm{x_i y_j P}(x_i,y_j)$$
$$=(1)(-4)(1/8)+(1)(2)(1/4)+(1)(7)(1/8)$$
$$+(5)(-4)(1/4)+(5)(2)(1/8)+(5)(7)(1/8)$$
$$=-\frac{1}{2}+\frac{1}{2}+\frac{7}{8}-5+\frac{5}{4}+\frac{35}{8}=\frac{3}{2}$$
$$\mathrm{COV(X,Y)}=\mathrm{E(XY)}-\mu_X\mu_Y$$
$$=(3/2)-(3)(1)=-\frac{3}{2}$$

(iii)

$$\sigma_X^2=\mathrm{E}\left(\mathrm{X^2}\right)-\mu_X^2 \text{ and}$$
$$\sigma_Y^2=\mathrm{E}\left(\mathrm{Y^2}\right)-\mu_Y^2$$

Now

$$\mathrm{E}\left(\mathrm{X^2}\right)=\sum x_i^2 \mathrm{p\,(x_i)}=(1)(1/2)+(25)(1/2)=13$$
$$\mathrm{E}\left(\mathrm{Y^2}\right)=\Sigma y_j^2 \mathrm{f\,(y_j)}=(16)(3/8)+(4)(3/8)+(49)(1/4)=\frac{79}{4}$$

Hence

$$\sigma_X^2 = 13 - (3)^2 = 4, \quad \sigma_Y^2 = (79/4) - (1)^2 = \frac{75}{4}$$

Thus $\sigma_X = 2$ and $\sigma_Y = \sqrt{\frac{75}{4}} = 4.33$

(iv)

$$\rho(\mathrm{X}, \mathrm{Y}) = \frac{COV(X,Y)}{\sigma_X \sigma_Y}$$

$$= \frac{-3/2}{(2)\sqrt{75/4}}$$

$$= \frac{-3}{2\sqrt{75}} = -0.1732$$

Thus $\mathrm{p}(\mathrm{X}, \mathrm{Y}) = -0.1732$

Problem 49. *The joint distribution of two random variables X and Y is as follows.*

$X \backslash Y$	1	3	9
2	$\frac{1}{8}$	$\frac{1}{24}$	$\frac{1}{12}$
4	$\frac{1}{4}$	$\frac{1}{4}$	0
6	$\frac{1}{8}$	$\frac{1}{24}$	$\frac{1}{12}$

Determine (i) Marginal Distributions of X and Y (ii) Covariance of X and Y [VTU Dec 2018, June 2011]

Solution : The distributions (marginal distribution) of X and Y are obtained by adding the all the respective row entries and also the respective column entries.

$$p(x=2) = \frac{1}{8} + \frac{1}{24} + \frac{1}{12} = \frac{1}{4}, \quad p(x=4) = \frac{1}{4} + \frac{1}{4} + 0 = \frac{1}{2}$$

$$p(x=6) = \frac{1}{8} + \frac{1}{24} + \frac{1}{12} = \frac{1}{4}, \quad p(y=1) = 1/8 + 1/4 + 1/8 = 1/2$$

$$p(y=3) = \frac{1}{24} + \frac{1}{4} + \frac{1}{24} = \frac{1}{3}, \quad p(y=9) = \frac{1}{12} + 0 + \frac{1}{12} = \frac{1}{6}$$

Hence individual(marginal) distributions of X and Y are,

x_i	2	4	6
$p(x_i)$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

y_j	1	3	9
$p(y_j)$	$\frac{1}{2}$	$\frac{1}{3}$	$\frac{1}{6}$

$$\mathrm{COV}(X,Y) = E(XY) - \mu_X \mu_Y$$

$$E(X) = \sum_i x_i\, p(x_i) = (2)(1/4) + (4)(1/2) + (6)(1/4) = 4$$

$$E(Y) = \sum_j y_j\, p(y_j) = (1)(1/2) + (3)(1/3) + (9)(1/6) = 3$$

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

$$= (2)(1)(1/8) + (2)(3)(1/24) + (2)(9)(1/12)$$
$$+ (4)(1)(1/4) + (4)(3)(1/4) + (4)(9)(0)$$
$$+ (6)(1)(1/8) + (6)(3)(1/24) + (6)(9)(1/12) = 12$$

$$\therefore \mathrm{COV(X,Y)} = E(XY) - \mu_X \mu_Y = 12 - (4)(3) = 0$$

Problem 50. *X and Y are independent random variables. X take values,* $2, 5, 7$ *with probability* $\frac{1}{2}, \frac{1}{4}$ *and* $\frac{1}{4}$*, respectively. Y takes values* $3, 4, 5$ *with the probability* $\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$

(a) Find the joint probability distribution of X and Y.

(b) Show that the covariance of X and Y is equal to zero.

(c) Find the probability distribution of Z = X + Y

Solution : Give data is as follows.

X_i	2	5	7
$p(xi)$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}$

and

y_j	3	4	5
$p(y_j)$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$

Given that X and Y are independent random variables.

Hence the joint probabilities are given by multiplying the individual probabilities.

i.e. $p(x_i, y_j) = p(x_i) \times p(y_j)$

Hence

$$p(x=2,y=3)=p(x=2)\times p(y=3)=\frac{1}{2}\times\frac{1}{3}=\frac{1}{6}$$

$$p(x=2,y=4)=p(x=2)\times p(y=4)=\frac{1}{2}\times\frac{1}{3}=\frac{1}{6}$$

$$p(x=2,y=5)=p(x=2)\times p(y=5)=\frac{1}{2}\times\frac{1}{3}=\frac{1}{6}$$

$$p(x=5,y=3)=p(x=5)\times p(y=3)=\frac{1}{4}\times\frac{1}{3}=\frac{1}{12}$$

$$p(x=5,y=4)=p(x=5)\times p(y=4)=\frac{1}{4}\times\frac{1}{3}=\frac{1}{12}$$

$$p(x=5,y=5)=p(x=5)\times p(y=5)=\frac{1}{4}\times\frac{1}{3}=\frac{1}{12}$$

$$p(x=7,y=3)=p(x=7)\times p(y=3)=\frac{1}{4}\times\frac{1}{3}=\frac{1}{12}$$

$$p(x=7,y=4)=p(x=7)\times p(y=4)=\frac{1}{4}\times\frac{1}{3}=\frac{1}{12}$$

$$p(x=7,y=5)=p(x=7)\times p(y=5)=\frac{1}{4}\times\frac{1}{3}=\frac{1}{12}$$

The joint distribution table is as follows.

X\Y	3	4	5	p (x_i)
2	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{2}$
5	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{4}$
7	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{4}$
p (y_j)	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	1

(b) $\mathbf{COV(X,Y) = E(XY) - \mu_X \mu_Y}$ where

$$\mu_X = \mathrm{E(X)} = \sum_i x_i p\left(x_i\right)$$
$$= (2)(1/2) + (5)(1/4) + (7)(1/4) = 4$$
$$\mu_Y = \mathrm{E(Y)} = \sum_i y_i p\left(y_i\right)$$
$$= (3)(1/3) + (4)(1/3) + (5)(1/3) = 4$$
$$\mathrm{E(XY)} = \sum_i x_i y_j P_{(x_i, y_j)}$$
$$= (2)(3)(1/6) + (2)(4)(1/6) + (2)(5)(1/6)$$
$$+ (5)(3)(1/12) + (5)(4)(1/12) + (5)(5)(1/12)$$
$$+ (7)(3)(1/12) + (7)(4)(1/12) + (7)(5)(1/12)$$
$$= 16$$

Hence

$$COV(\mathrm{X,Y}) = \mathrm{E(XY)} - \mu_X \mu_Y = 16 - (4)(4) = 0$$

(c) Let $Z = X + Y$

i.e. $z_i = x_i + y_i$ and hence $\{z_i\} = \{5, 6, 7, 8, 9, 10, 11, 12\}$

The corresponding probabilities are,

Z	5	6	7	8	9	10	11	12
$P(Z)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{6}$	$\frac{1}{12}$	$\frac{1}{12}$

Problem 51. *X and Y are independent random variables. X takes the values $1, 2$ with probability $0.7, 0.3$ each and y takes the values $-2, 5, 8$ with probabilities $0.3, 0.5, 0.2$. Find the joint distribution of X and Y. Hence find $COV(X, Y)$. [VTU Jan 2018]*

Solution : Given distributions are

x_i	1	2
$p\left(x_i\right)$	0.7	0.3

y_j	-2	5	8
$p\left(y_j\right)$	0.3	0.5	0.2

Since X and Y are independent, the joint probabilities $p(x_i, y_j)$ is obtained by using $p(x_i, y_j) = p\left(x_i\right) \times p\left(y_j\right)$

$p(1,-2) = (0.7)(0.3) = 0.21$, $p(1,5) = (0.7)(0.5) = 0.35$
$p(1,8) = (0,7)(0.2) = 0.14$, $p(2,-2) = (0.3)(0.3) = 0.09$,
$p(2,5) = (0.3)(0.5) = 0.15$, $p(2,8) = (0.3)(0.2) = 0.06$

Hence the joint distribution table is,

$X \backslash Y$	-2	5	8	$p(x_i)$
1	0.21	0.35	0.14	0.7
2	0.09	0.15	0.06	0 .3
$p(y_i)$	0.3	0.5	0.2	1

We have $\mathbf{COV}(X,Y) = E(XY) - \mu_X \mu_Y$ where

$$\mu_X = E(X) = \sum_i x_i p(x_i) = (1)(0.7) + (2)(0.3) = 1.3$$

$$\mu_Y = E(Y) = \sum_j y_j p(y_j) = (-2)(0.3) + (5)(0.5) + (8)(0.2) = 3.5$$

$$E(XY) = \sum_{i,j} x_i y_j P_{(x_i,y_j)}$$
$$= (1)(-2)(0.21) + (1)(5)(0.35) + (1)(8)(0.14)$$
$$+ 2(-2)(0.09) + (2)(5)(0.15) + (2)(8)(0.06)$$
$$= -0.42 + 1.75 + 1.12 - 0.36 + 1.5 + 0.96 = 4.55$$

Hence

$\mathbf{COV}(X,Y) = 4.55 - (1.3)(3.5) = 0$

Thus the result $\mathbf{COV}(X,Y) = 0$ for independent random variables X and Y is verified.

Problem 52. *The joint distribution of two discrete random variables X and Y is* $f(x,y) = k(2x+y)$ *where x and y are integers such that* $0 \le x \le 2$, $0 \le y \le 3$. *Find (i) The value of k (ii) Marginal distributions of X and Y (iii) Are X and Y independent ? (iv)Find* $P(X \ge 1; Y \le 2)$ *(v)* $P(X+Y > 2)$ *[VTU Jan 2018]*

Solution : Given that x and y are integers such that $0 \le x \le 2$, $0 \le y \le 3$

Hence values of X and Y can be listed as

$X = 0, 1, 2$ and $Y = 0, 1, 2, 3$

Using these x and y values we can calculate $f(x,y) = k(2x+y)$, in terms of unknown k, and the joint probability distribution table is formed as follows.

$X \backslash Y$	0	1	2	3	Sum
0	0	k	$2k$	$3k$	$6k$
1	$2k$	$3k$	$4k$	$5k$	$14k$
2	$4k$	$5k$	$6k$	$7k$	$22k$
Sum	$6k$	$9k$	$12k$	$15k$	$\boxed{42k}$

(a) From the definition of joint pdf, we must have

$\sum_{i,j} p(x_i, y_j) = 1 \Rightarrow 42k = 1 \quad \therefore \quad k = \frac{1}{42}$

(b) Marginal probability distribution of X and Y are given by

x_i	0	1	2
$p(x_i)$	$\frac{1}{7}$	$\frac{1}{3}$	$\frac{11}{21}$

y_j	0	1	2	3
$p(y_j)$	$\frac{1}{7}$	$\frac{3}{14}$	$\frac{2}{7}$	$\frac{5}{14}$

(c)X and Y are independent if they satisfy the condition

$$p(x_i, y_j) = p(x_i) \times p(y_j)$$

for all $x_i \in X$ and $y_j \in Y$.

From the joint table, we have $p(X = 0, Y = 0) = 0$

From individual Tables, we have

$p(X = 0) \times p(Y = 0) = \frac{1}{7} \times \frac{1}{7} = \frac{1}{49}$

Clearly, $p(x_i)\, p(y_j) \neq P(x_i, y_j)$

Hence the random variables are dependent.

In other words, X and Y are not independent.

The joint probabilities favourable to $(X \geq 1, Y \leq 2)$ are corresponding to (X, Y) with values $(1,0), (1,1), (1,2), (2,0), (2,1), (2,2)$

Adding the corresponding probabilities, we get

$$\begin{aligned} P(X \geq 1, Y \leq 2) &= p(1,0) + p(1,1) + p(1,2) \\ &\quad + p(2,0) + p(2,1) + p(2,2) \\ &= \frac{2}{42} + \frac{3}{42} + \frac{4}{42} + \frac{4}{42} + \frac{5}{42} + \frac{6}{42} \\ &= \frac{24}{42} = \frac{4}{7} \end{aligned}$$

To have $X + Y > 2$ the values of (X, Y) are,
$(0,3), (1,2), (1,3), (2,1), (2,2), (2,3)$

$$P(X+Y>2) = p(0,3) + p(1,2) + p(1,3) + p(2,1) + p(2,2) + p(2,3)$$
$$= \frac{3}{42} + \frac{4}{42} + \frac{5}{42} + \frac{5}{42} + \frac{6}{42} + \frac{7}{42}$$
$$= \frac{30}{42} = \frac{5}{7}$$

Problem 53. *A fair coin is tossed thrice. The random variables X and Y are defined as follows:* $X = 0, 1$ *according as head or tail occurs on the first toss, y= no. of heads.*
(i) Determine the marginal probability distribution of X and Y (ii)Determine the joint distribution of X and Y (iii) Determine E(X), E(Y) and E(XY) (iv) Determine σ_x, σ_y *[VTU]*

Solution :
The sample space S and the association of random variables X and Y is given by the following table.

S	HHH	HHT	HTH	HTT	THH	THT	TTH	TTT
X	0	0	0	0	1	1	1	1
Y	3	2	2	1	2	1	1	0

Here $X = 0, 1$ and $Y = 0, 1, 2, 3$

$$P(X=0) = \frac{4}{8} = \frac{1}{2}, \quad P(X=1) = \frac{4}{8} = \frac{1}{2}$$
$$P(Y=0) = \frac{1}{8}, \quad P(Y=1) = \frac{3}{8}$$
$$P(Y=2) = \frac{3}{8}, \quad P(Y=3) = \frac{1}{8}$$

Thus we have the following probability distribution of X and Y as :

x_i	0	1	y_j	0	1	2	3
$p(x_i)$	$\frac{1}{2}$	$\frac{1}{2}$	$p(y_i)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

(b) The joint distribution of X and Y is found by computing $P(x_i, y_j) = P(X = x_i, Y =$ where we have $x = 0, 1$ and $y = 0, 1, 2, 3$

$P(X = 0, Y = 0) = 0$

($X = 0$ implies that there is a head turn out and $Y = 0$ implies that there the total number he
which is impossible event).

$P(X = 0, Y = 1) = \frac{1}{8}$ (corresponding to the outcome HTT)

$P(X = 0, Y = 2) = \frac{2}{8} = \frac{1}{4}$ (out comes are HHT and HTH)

$P(X = 0, Y = 3) = \frac{1}{8}$; (outcome is HHH)

$P(X = 1, Y = 0) = \frac{1}{8}$; (outcome is TTT)

$P(X = 1, Y = 1) = \frac{2}{8} = \frac{1}{4}$; (out comes are THT, TTH)

$P(X, = 1, Y = 2) = \frac{1}{8}$; (outcome is THH)

$P(X = 1, Y = 3) = 0$; (since the outcome is impossible).

(The above values can be written by looking at the table of S, X, Y)

The required joint probability distribution of X and Y is as follows.

$X \backslash Y$	0	1	2	3	Sum of row entries
0	0	$\frac{1}{8}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{2}$
1	$\frac{1}{8}$	$\frac{1}{4}$	$\frac{1}{8}$	0	$\frac{1}{2}$
Sum of column entries	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$	1

$$\mu_X = E(X) = \sum x_i p(x_i)$$
$$= (0)(1/2) + (1)(1/2) = 1/2$$
$$\mu_Y = E(Y) = \sum_i y_j p(y_j)$$
$$= (0)(1/8) + (1)(3/8) + (2)(3/8) + (3)(1/8)$$
$$= 12/8 = 3/2$$
$$E(XY) = \sum_{i,j} x_i y_j P(x_i, y_j)$$
$$= (0)(0)(0) + (0)(1)\frac{1}{8} + (0)(2)\frac{1}{4} + (0)(3)\frac{1}{8}$$
$$+ (1)(0)\frac{1}{8} + (1)(1)\frac{1}{4} + (1)(2)\frac{1}{8} + (1)(3)(0)$$
$$= \frac{1}{2}$$

$$\sigma_X^2 = E\left(X^2\right) - \mu_X^2 = \sum x_i^2 p\left(x_i\right) - \mu_X^2$$
$$= (0)^2(1/2) + (1)^2(1/2) - 1/4 = 1/4$$
$$\Rightarrow \sigma_X = \frac{1}{2}$$
$$\sigma_Y^2 = E\left(Y^2\right) - \mu_Y^2$$
$$= (0)^2(1/8) + (1)^2(3/8) + (2)^2(3/8) + (3)^2(1/8) - (9/4)$$
$$= 3/4$$
$$\Rightarrow \sigma_Y = \frac{\sqrt{3}}{2}$$

$$\mathrm{COV}(X,Y) = E(XY) - \mu_X\mu_Y = \frac{1}{2} - \frac{2}{4} = -\frac{1}{4}$$
$$\rho(X,Y) = \frac{\mathrm{COV}(X,Y)}{\sigma_X\sigma_Y} = \frac{-\frac{1}{4}}{\frac{\sqrt{3}}{4}} = -\frac{1}{\sqrt{3}}$$

Problem 54. *The joint distribution of two random variables X and Y is as follows.*

X / Y	*-2*	*-1*	*4*	*5*
1	*0.1*	*0.2*	*0*	*0.3*
2	*0.2*	*0.1*	*0.1*	*0*

Determine (i) Marginal Distributions of X and Y (ii) Covariance of X and Y (iii) Correlation of X and Y (iv) Find $p(X+Y>0)$ *[VTU Jan 2020,June 2019, Dec 2010]*

Solution : The marginal distribution of X is given by

X	1	2
$p_X(x)$	0.6	0.4

The marginal distribution of Y is given by

Y	-2	-1	4	5
$p_Y(y)$	0.3	0.3	0.1	0.3

(i)

$$E(X) = \sum_x x f_X(x) = 1 \times 0.6 + 2 \times 0.4 = 1.4$$
$$E(Y) = \sum_y y f_Y(y) = (-2)(0.3) + (-1)(0.3) + 4(0.1) + 5(0.3) = 1$$

(ii)

$$E(XY) = \sum_x \sum_y xyp(x,y)$$
$$= 1 \times (-2) \times 0.1 + 1 \times (-1) \times 0.2 + 1 \times 4 \times 0$$
$$+ 1 \times 5 \times 0.3 + 2 \times (-2) \times 0.2 + 2 \times (-1) \times 0.1$$
$$+ 2 \times 4 \times 0.1 + 2 \times 5 \times 0 = 0.9$$
$$\therefore COV\ (X,Y) = E(XY) - \mu_X \mu_Y = 0.9 - 1.4 \times 1 = -0.5$$

(iii)

$$E\left(X^2\right) = \sum_x x^2 p_x(x) = 1^2 \times 0.6 + 2^2 \times 0.4 = 2.2$$
$$\therefore\ \sigma_X^2 = E\left(X^2\right) - \mu_x^2 = 2.2 - (1.4)^2 = 0.24$$
$$\Rightarrow \sigma_x = 0.49$$
$$E\left(Y^2\right) = \sum_y y^2 f_Y(y)$$
$$= (-2)^2 \times 0.3 + \left(1^2\right) \times 0.3 + 4^2 \times 0.1 + 5^2 \times 0.3 = 10.6$$
$$\therefore\ \sigma_Y^2 = E\left(Y^2\right) - \mu_Y^2 = 10.6 - 1^2 = 9.6$$
$$\Rightarrow\ \sigma_Y = 3.1$$

Now,

$$\rho(X,Y) = \frac{\text{Cov}(X,Y)}{\sigma_X \sigma_Y}$$
$$= \frac{-0.5}{0.49 \times 3.1} = -0.329$$

(v) $X+Y > 0$ is possible when (X,Y) take the values $(1,4)$; $(1,5)$; $(2,-1)$; $(2,4)$ and $(2,5)$

Hence

$$P(X+Y>0) = p(1,4) + p(1,5) + p(2,-1) + p(2,4) + p(2,5)$$
$$= 0 + 0.3 + 0.1 + 0.1 + 0$$
$$= 0.5$$

Problem 55. *A fair coin is tossed 4 times. Let X denote the no. of heads occurring and let Y denote the longest string of heads occurring. Find (i) the joint distribution of X and Y (ii) Marginal distributions of X and Y (iii)COV(X,Y) [VTU June 2011]*

Solution : When a fair coin is tossed four times, the set of all possible outcomes

$$S = \{TTTT, THTT, HTTT, TTHT, TTTH, TTHH, HTTH, THHT,$$
$$THTH, HHTT, HTHT, THHH, HTHH, HHTH, HHHT, HHHH\}$$

There are $2^4 = 16$ possible outcomes.

X is the number of heads and it can take values $0, 1, 2, 3, 4$ and Y is the length of the longest string of heads that can take values $0, 1, 2, 3, 4$

We calculate the joint probabilities as follows.

$f(0,0) = P(TTTT) = \frac{1}{16}$

$f(0,1) = f(0,2) = f(0,3) = f(0,4) = 0, f(1,0) = 0$

$f(1,1) = P[THTT, HTTT, TTHT, TTTH] = \frac{4}{16}$

$f(2,0) = 0$

$f(2,1) = P[THTH, HTHT, HTTH] = \frac{3}{16}$

$f(2,2) = P[TTHH, THHT, HHTT] = \frac{3}{16}$

$f(2,3) = f(2,4) = 0, f(3,0) = f(3,1) = 0$

$f(3,2) = P[HTHH, HHTH] = \frac{2}{16}$

$f(3,3) = P[THHH, HHHT] = \frac{2}{16}$

$f(3,4) = 0$

$f(4,0) = f(4,1) = f(4,2) = f(4,3) = 0$

$f(4,4) = P[HHHH] = \frac{1}{16}$

The joint distribution table is,

X/Y	0	1	2	3	4	Row Sum
0	$\frac{1}{16}$	0	0	0	0	$\frac{1}{16}$
1	0	$\frac{4}{16}$	0	0	0	$\frac{4}{16}$
2	0	$\frac{3}{16}$	$\frac{3}{16}$	0	0	$\frac{6}{16}$
3	0	0	$\frac{2}{16}$	$\frac{2}{16}$	0	$\frac{4}{16}$
4	0	0	0	0	$\frac{1}{16}$	$\frac{1}{16}$
Column Sum	$\frac{1}{16}$	$\frac{7}{16}$	$\frac{5}{16}$	$\frac{2}{16}$	$\frac{1}{16}$	①

(b) The marginal distribution of X is

X	0	1	2	3	4
$f_X(x)$	$\frac{1}{16}$	$\frac{4}{16}$	$\frac{6}{16}$	$\frac{4}{16}$	$\frac{1}{16}$

The marginal distribution of Y is given by

Y	0	1	2	3	4
$f_Y(y)$	$\frac{1}{16}$	$\frac{7}{16}$	$\frac{5}{16}$	$\frac{2}{16}$	$\frac{1}{16}$

$$\mu_x = E(X) = \sum_x x f_X(x)$$
$$= 0 \times \frac{1}{16} + 1 \times \frac{4}{16} + 2 \times \frac{6}{16} + 3 \times \frac{4}{16} + 4 \times \frac{1}{16}$$
$$= 2$$
$$\mu_Y = E(Y) = \sum_y y f_Y(y)$$
$$= 0 \times \frac{1}{16} + 1 \times \frac{7}{16} + 2 \times \frac{5}{16} + 3 \times \frac{2}{16} + 4 \times \frac{1}{16}$$
$$= \frac{27}{16}$$
$$\text{Now, } E(XY) = \sum_x \sum_y xy f(x,y)$$
$$= 0 \times 0 \times \frac{1}{16} + 1 \times 1 \times \frac{4}{16}$$
$$+ 2 \times 1 \times \frac{3}{16} + 2 \times 2 \times \frac{3}{16} +$$
$$+ 3 \times 2 \times \frac{2}{16} + 3 \times 3 \times \frac{2}{16} + 4 \times 4 \times \frac{1}{16}$$
$$= \frac{17}{4}$$
$$\therefore \text{Cov}(X,Y) = E(XY) - \mu_X \mu_Y = \frac{17}{4} - 2 \times \frac{27}{16} = 0.875$$

2.4 Stochastic process

Let S be the sample space, representing the set of all possible outcomes of a random experiment, and R be the set of all real numbers. A random variable X is a function f from S to R, denoted as $X = f(s)$, where $s \in S$. We introduce an index set $T \subset R$, indexed by the parameter t representing time.

Suppose the value of a random variable defined on S depends on $s \in S$ and $t \in T$. In this context, a stochastic process is a collection of random variables $\{X(t), t \in T\}$, where each $X(t)$ represents the random variable associated with time t. The index set T could be discrete or continuous, representing time or some other parameter.

Here, $X_0 = X(0)$ is the value of the stochastic process at the initial time point, referred to as the initial state of the system.
Example : Let's consider a simple example of the daily temperature in a city. Each day, the temperature can be considered a random variable, and the entire sequence of daily temperatures forms a stochastic process.
Let's denote the daily temperature as $X(t)$, where t is the day of the year. The index set T is the set of all days in a year. Mathematically, we can represent this stochastic process as $\{X(t), t \in T\}$, where each $X(t)$ is the temperature on day t. The initial state of the system, X_0, corresponds to the temperature on the first day of the year.

This stochastic process captures the idea that daily temperatures vary randomly throughout the year. The randomness is introduced by factors such as weather patterns, seasonality, and other atmospheric conditions.
Classification of Stochastic Processes :

2.5 Vector:

A vector is a tuple of numbers $(v_1, v_2, \ldots, v_n)$ where the quantities $v_1, v_2, \ldots, v_n$ are called components of the vector.

2.6 Probability Vector:

A probability vector is a vector in which each component represents the probability of an event. In other words, let $v = [v_1, v_2, \ldots, v_n]$ be a vector, where each v_i is a non-negative number representing the probability of event i, and the sum of all components equals 1: $\sum_{i=1}^{n} v_i = 1$, then v is called a probability vector.
Examples: $\mathrm{u} = (1, 0)$; $\mathrm{v} = \left(\frac{1}{2}, \frac{1}{2}\right)$; $\mathrm{w} = \left(\frac{1}{4}, \frac{1}{4}, \frac{1}{2}\right)$ are all probability vectors.
Which vectors are probability vectors?
(a) $\mathrm{u} = \left(\frac{1}{4}, \frac{1}{2}, -\frac{1}{4}, \frac{1}{2}\right)$
(b) $\mathrm{v} = \left(\frac{1}{2}, 0, \frac{1}{3}, \frac{1}{6}, \frac{1}{6}\right)$
(c) $\mathrm{w} = \left(\frac{1}{12}, \frac{1}{2}, \frac{1}{6}, \frac{1}{4}\right)$
(d) $\mathrm{r} = \left(\frac{1}{3}, 0, 0, \frac{1}{6}, \frac{1}{2}\right)$

Here, (a) $\mathbf{u}$ is not a probability vector since its third component is negative.
(b) $\mathbf{v}$ is not a probability vector since the sum of the components is not equal to 1 .
(c) $\mathbf{w}$ is a probability vector since the components are non negative and their sum is 1 .
(d) $\mathbf{r}$ is also a probability vector since the components are non negative and their sum is 1 .

2.7 Stochastic Matrix:

A stochastic matrix is a square matrix in which each entry is a non-negative real number, and the sum of the entries in each row is equal to 1. Let $\mathrm{P} = [p_{ij}]$ be an $n \times n$ matrix. It is stochastic if for each j, $\sum_{i=1}^{n} p_{ij} = 1$, and all $p_{ij} \geq 0$.

Example of a 2 by 2 Stochastic Matrix:

$$\mathrm{P} = \begin{bmatrix} 0.6 & 0.4 \\ 0.3 & 0.7 \end{bmatrix}$$

Example of a 3 by 3 Stochastic Matrix:

$$\mathrm{P} = \begin{bmatrix} 0.2 & 0.5 & 0.3 \\ 0.4 & 0.1 & 0.5 \\ 0.3 & 0.4 & 0.3 \end{bmatrix}$$

Problem 56. *Which matrices are stochastic?*

(a) $\boldsymbol{P} = \begin{bmatrix} 0 & 1 & 0 \\ \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \end{bmatrix}$,

(b) $\boldsymbol{Q} = \begin{bmatrix} 0 & 1 \\ \frac{1}{2} & \frac{1}{4} \end{bmatrix}$,

(c) $\boldsymbol{R} = \begin{bmatrix} \frac{1}{4} & \frac{3}{4} \\ \frac{1}{6} & \frac{5}{6} \end{bmatrix}$

Solution : Here, (a) $\boldsymbol{P}$ is not a square matrix. So it is not stochastic.
(b) The sum of elements in the second row is not equal to 1 . So $\boldsymbol{Q}$ is not stochastic.
(c) $\boldsymbol{R}$ is a stochastic matrix since it is a square matrix with all elements non negative and sum of elements in each row is equal to 1 .

2.8 Regular stochastic matrix :

A regular stochastic matrix is a stochastic matrix that has a power (positive integer power) with all positive entries. Specifically, a stochastic matrix $\mathbf{P}$ is regular if there exists a positive integer $\boldsymbol{k}$ such that $\mathbf{P}^k$ has all positive entries, different from zero.
Note : A stochastic matrix P is not regular if a 1 occurs in the principal main diagonal.

Example: $\boldsymbol{A} = \begin{bmatrix} 0 & 1 \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix}$ Then $\boldsymbol{A^2} = \begin{bmatrix} 0 & 1 \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix}\begin{bmatrix} 0 & 1 \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{4} & \frac{3}{4} \end{bmatrix}$ $\therefore$ $\boldsymbol{A}$ is a regular stochastic matrix $(\boldsymbol{k = 2})$

(b)Consider $\boldsymbol{A} = \begin{bmatrix} 0 & 0 & 1 \\ \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 1 & 0 \end{bmatrix}$

Here,

$$\boldsymbol{A^2} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} \end{bmatrix}$$

,

$$\boldsymbol{A^3} = \begin{bmatrix} \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \\ 0 & \frac{1}{2} & \frac{1}{2} \end{bmatrix}$$

,

$$\boldsymbol{A^4} = \begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{2} \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \end{bmatrix}$$

,

$$\boldsymbol{A^5} = \begin{bmatrix} \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \\ \frac{1}{8} & \frac{1}{2} & \frac{3}{8} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{2} \end{bmatrix}.$$

All elements in $\boldsymbol{A^5}$ are positive. Hence $\boldsymbol{A}$ is regular.

2.9 Properties of a Regular Stochastic Matrix :

The following properties are associated with a regular stochastic matrix $\boldsymbol{P}$ of order $\boldsymbol{n}$.

(i) $\boldsymbol{P}$ has a unique fixed point $\boldsymbol{x}$ such that $\boldsymbol{xP = x}$.

(ii) $\boldsymbol{P}$ has a unique fixed probability vector $\boldsymbol{v}$ such that $\boldsymbol{vP} = \boldsymbol{v}$ and $\sum \boldsymbol{v_i} = \mathbf{1}$.

(iii) $\boldsymbol{P^2}, \boldsymbol{P^3}, \ldots$ approaches the matrix $\boldsymbol{V}$ whose rows are each the fixed probability vector $\boldsymbol{v}$.

(iv) If $\boldsymbol{u}$ is any probability vector, then the sequence of vectors $\boldsymbol{uP}, \boldsymbol{uP^2}, \ldots$ approaches the unique fixed probability vector $\boldsymbol{v}$.

Problem 57. *Find the unique fixed probability vector for the regular stochastic matrix*

$$\boldsymbol{A} = \begin{bmatrix} 0 & 1 & 0 \\ 1/6 & 1/2 & 1/3 \\ 0 & 2/3 & 1/3 \end{bmatrix}$$

Solution : We have to find $\boldsymbol{v} = (v_1, v_2, v_3)$ where $v_1 + v_2 + v_3 = 1$ such that $\boldsymbol{vA} = \boldsymbol{v}$:

$$\begin{bmatrix} v_1 & v_2 & v_3 \end{bmatrix} \begin{bmatrix} 0 & 1 & 0 \\ \frac{1}{6} & \frac{1}{2} & \frac{1}{3} \\ 0 & \frac{2}{3} & \frac{1}{3} \end{bmatrix} = \begin{bmatrix} v_1 & v_2 & v_3 \end{bmatrix}$$

$$\left[\frac{v_2}{6}v_1 + \frac{v_2}{2} + \frac{2v_3}{3}, \frac{v_2}{3} + \frac{v_3}{3} - \right] = [v_1, v_2, v_3]$$

This leads to the system of equations:

$$\frac{v_2}{6} = v_1$$
$$v_1 + \frac{v_2}{2} + \frac{2z}{3} = v_2$$
$$\frac{v_2}{3} + \frac{v_3}{3} = v_3$$
$$v_1 + v_2 + v_3 = 1$$

This gives,

$$v_2 = 6v_1$$
$$6v_1 + 3v_2 + 4v_3 = 6v_2$$
$$v_2 - 2v_3 = 0v_1 + v_2 + v_3 = 1$$
$$-6v_1 + v_2 = 0 \qquad (1)$$
$$6v_1 - 3v_2 + 4v_3 = 0 \qquad (2)$$
$$0v_1 + v_2 - 2v_3 = 0 \qquad (3)$$
$$v_1 + v_2 + v_3 = 1 \qquad (4)$$

When solving the system of equations, please note that all four equations (equations (1), (2), (3), and (4)) are essential for obtaining a meaningful solution. Equation (4)

provides the information required for a probability vector and ensures a unique and non-trivial solution.

Please select any two equations from (1), (2), and (3) and solve them together with (4).

For example, by solving (1), (2), and (4), we find the solution:

$$v_1 = \frac{1}{10}, \quad v_2 = \frac{6}{10}, \quad v_3 = \frac{3}{10}$$

To confirm the accuracy of this solution, substitute these values into equation (3) and check if it holds true. Thus, the required unique fixed probability vector v is given by $v = (v_1, v_2, v_3) = (\frac{1}{10}, \frac{6}{10}, \frac{3}{10})$.

Problem 58. *Show that* $P = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ \frac{1}{2} & \frac{1}{2} & 0 \end{bmatrix}$ *ia a regular stochastic matrix and find the corresponding unique fixed probability vector.*

Solution: P is a regular stochastic matrix if there exists an integer $n \geq 1$ such that all entries in P^n are ¿0.

Hence let us find the powers of P.

$$P = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ \frac{1}{2} & \frac{1}{2} & 0 \end{bmatrix}$$

$$\text{Consider } P^2 = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ \frac{1}{2} & \frac{1}{2} & 0 \end{bmatrix} \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ \frac{1}{2} & \frac{1}{2} & 0 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 1 \\ \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & \frac{1}{2} \end{bmatrix}$$

$$P^3 = P \cdot P^2 = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ \frac{1}{2} & \frac{1}{2} & 0 \end{bmatrix} \cdot \begin{bmatrix} 0 & 0 & 1 \\ \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & \frac{1}{2} \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{4} & \frac{1}{4} & 1/2 \end{bmatrix}$$

$$P \cdot P^3 = P^4 = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ \frac{1}{2} & \frac{1}{2} & 0 \end{bmatrix} \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{2} \end{bmatrix} = \begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{2} \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \end{bmatrix}$$

$$P^5 = P \cdot P^4 = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ \frac{1}{2} & \frac{1}{2} & 0 \end{bmatrix} \begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{2} \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \end{bmatrix} = \begin{bmatrix} \frac{1}{4} & \frac{1}{4} & \frac{1}{2} \\ v & \frac{1}{2} & \frac{1}{4} \\ \frac{1}{8} & \frac{3}{8} & \frac{1}{2} \end{bmatrix}$$

We observe that in P^5 all the entries are positive.

Thus P is a regular stochastic matrix.

Next we have to find $v = (v_1, v_2, v_3)$ where $v_1 + v_2 + v_3 = 1$ such that $vP = v$

$$\Rightarrow \quad [v_1, v_2, v_3] \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1/2 & 1/2 & 0 \end{bmatrix} = [v_1, v_2, v_3]$$

$$\text{i.e.,} \quad \left[\frac{v_3}{2}, v_1 + \frac{v_3}{2}, v_2\right] = [v_1, v_2, v_3]$$

$$\Rightarrow \quad \frac{v_3}{2} = v_1, v_1 + \frac{v_3}{2} = v_2, v_2 = v_3$$

Rearranging, we get

$$-v_1 + 0v_2 + \frac{v_3}{2} = 0 \qquad (1)$$

$$v_1 - v_2 + \frac{v_3}{2} = 0 \qquad (2)$$

$$0v_1 + v_2 - v_3 = 0 \qquad (3)$$

$$v_1 + v_2 + v_3 = 1 \qquad (4)$$

by solving (1), (2), and (4), we find the solution:

$$v_1 = \frac{1}{5}, \quad v_2 = \frac{2}{5}, \quad v_3 = \frac{2}{5}$$

Thus $[v_1, v_2, v_3] = \left[\frac{1}{5}, \frac{2}{5}, \frac{2}{5}\right]$ is the required unique fixed probability vector of P.

Problem 59. *Find the unique fixed probability vector for the regular stochastic matrix,*

$$A = \begin{bmatrix} 0 & \frac{3}{4} & \frac{1}{4} \\ \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & 1 & 0 \end{bmatrix}$$

Solution : Given that

$$A = \begin{bmatrix} 0 & \frac{3}{4} & \frac{1}{4} \\ \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & 1 & 0 \end{bmatrix}$$

is a regular stochastic matrix.

We have to find $v = (v_1, v_2, v_3)$ such that $vA = v$ and $v_1 + v_2 + v_3 = 1$

i.e.

$$\begin{bmatrix} v_1 & v_2 & v_3 \end{bmatrix} \begin{bmatrix} 0 & \frac{3}{4} & \frac{1}{4} \\ \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & 1 & 0 \end{bmatrix} = \begin{bmatrix} v_1 & v_2 & v_3 \end{bmatrix}$$

and $v_1 + v_2 + v_3 = 1$ This leads to the system of equations:

$$0v_1 + \frac{3}{4}v_2 + \frac{1}{4}v_3 = v_1$$
$$\frac{1}{2}v_1 + \frac{1}{2}v_2 + 0v_3 = v_2$$
$$0v_1 + 1v_2 + 0v_3 = v_3$$
$$v_1 + v_2 + v_3 = 1$$

Rearranging, we get

$$-v_1 + \frac{3}{4}v_2 + \frac{1}{4}v_3 = 0 \quad (1)$$
$$\frac{1}{2}v_1 - \frac{1}{2}v_2 + 0v_3 = 0 \quad (2)$$
$$0v_1 + 1v_2 - v_3 = 0 \quad (3)$$
$$v_1 + v_2 + v_3 = 1 \quad (4)$$

When solving the system of equations, please note that all four equations (equations (1), (2), (3), and (4)) are essential for obtaining a meaningful solution. Equation (4) provides the information required for a probability vector and ensures a unique and non-trivial solution.

Select any two equations from (1), (2), and (3) and solve them together with (4).

For example, by solving (2), (3), and (4), we find the solution:

$v_1 = \frac{1}{4}$, $v_2 = \frac{1}{2}$, and $v_3 = \frac{1}{2}$.

To confirm the accuracy of this solution, substitute these values into equation (1) and check if it holds true. Thus, the required unique fixed probability vector v is given by

$$v = \left(\frac{1}{4}, \frac{1}{2}, \frac{1}{2}\right)$$

.

2.10 State and State space

Stochastic Process is a family of random variables $\{X(t) \mid t \in T\}$ defined on a common sample space S and indexed by the parameter t, which varies on an index set T. The values assumed by the random variables $X(t)$ are called **states**, and the set of all possible values from the **state space** of the process is denoted by I. If the state space is discrete, the stochastic process is known as a **chain**. In this case the state space is assumed to be $I = \{a_1, a_2, \ldots\}$. Thus a (finite) stochastic process

consists of a sequence of experiments in which each experiment has a finite number of outcomes with given probabilities.

2.11 Markov chain

Consider a stochastic process $\{X_0, X_1, X_2, \ldots\}$ having the state space $I = \{a_1, a_2, \ldots a_m\}$. i.e. $X_0, X_1, X_2, \ldots$ is a sequence of random variables and $a_1, a_2, \ldots a_m$ are the possible values of these random variables. When $X_n = a_i$, we say that the system is in state a_i at time n, or at the n^{th} step. If the probability that $X_{n+1} = a_j$ given $X_n = a_i$ is independent of the states $X_0, X_1, X_2, \ldots, X_{n-1}$ then we say that the stochastic process $\{X_0, X_1, X_2, \ldots\}$ is a **Markov(memoryless) Chain**. i.e., in a Markov Chain the future outcomes depend only on the present outcome and is independent of the past outcomes.

2.12 Transition probabilities

In a Markov chain, the system undergoes transitions from one state to another, and the probability of transitioning to any particular state depends solely on the current state, not on the sequence of events that preceded it. This property is known as the Markov property. Associated with each ordered pair of states (a_i, a_j), the number p_{ij} gives the probability that system changes from i th state to j th state. i.e. Let $p_{ij} = P\{X_{n+1} = a_j \mid X_n = a_i\}$. In other words, p_{ij} is the probability that a_j occurs immediately after a_i occurs. The numbers p_{ij} are known as **transition probabilities**.

The transition probabilities p_{ij} satisfy $p_{ij} \geq 0$ and $\sum_{j=1}^{m} p_{ij} = 1$, for $i = 1, 2, \ldots, m$

i.e. when $i = 1$, we have $p_{11} + p_{12} + \cdots p_{1m} = 1$

when $i = 2$, we have $p_{21} + p_{22} + \cdots p_{2m} = 1$

$\vdots$

when $i = m$, we have $p_{m1} + p_{m2} + \cdots p_{mm} = 1$

These probabilities may be arranged in the matrix form, P which is the square matrix

of the transition probabilities p_{ij} in the form:

$$P = \begin{array}{c} \\ (a_1) \\ (a_2) \\ \vdots \\ (a_m) \end{array} \begin{array}{c} \begin{array}{ccccc} (a_1) & (a_2) & (a_3) & \dots & (a_m) \end{array} \\ \begin{pmatrix} p_{11} & p_{12} & p_{13} & \dots & p_{1m} \\ p_{21} & p_{22} & p_{23} & \dots & p_{2m} \\ \dots & \dots & \dots & \dots & \dots \\ p_{m1} & p_{m2} & p_{m3} & \dots & p_{mm} \end{pmatrix} \end{array}.$$

Here, the i th row of P namely $(p_{i1}, p_{i2}, \dots, p_{im})$ represents the probabilities of that system will change from a_i to $a_1, a_2, a_3, \dots, a_m$ respectively. P is called the **transition probability matrix (t.p.m.)** of the Markov Chain.

Note that **each row of P is a probability vector** and so P is a stochastic matrix.(i.e. in each row of P sum of the entries is always equal to 1)

2.13 Irreducible Markov Chains

A Markov Chain is said to be **irreducible** if every state can be reached from every other state, i.e., for any two states a_i and a_j, we have $p_{ij}^{(n)} > 0$ for some $n \geq 1$. When the transition matrix of the Markov Chain is regular, all elements of some power of P are positive. Hence the Markov Chain is irreducible when P is regular.

2.14 n-step transition probabilities

If P is the transition matrix of a Markov chain and if the probability that a Markov chain will move from state i to state j in exactly n steps is denoted by $p_{ij}(n)$ or $p_{ij}^{(n)}$ then the n-step transition matrix $P^{(n)}$ is equal to the nth power of P, i.e., $P^{(n)} = P^n$.

In other words, the problem of finding the n-step transition probabilities is reduced to one of forming powers of a given matrix.

Probability distribution of the system at some arbitrary time is denoted by the probability vector.

$$p = (p_1, p_2, \dots, p_m) = (p(a_1), p(a_2), p(a_3), \dots, p(a_m))$$

where $p_i = p(a_i)$ denotes the probability that the system is in state a_i.
At time $t = 0$(i.e. initial state), when the process begins, the corresponding probability vector

$$p^{(0)} = \left(p_1^{(0)}, p_2^{(0)}, \ldots, p_i^{(0)}, \ldots, p_m^{(0)}\right)$$

denotes the initial probability distribution. Then the probability distribution of the system n-steps later, for $n = 1, 2, 3 \cdots$ is given by

$$p^{(1)} = p^{(0)}P,$$
$$p^{(2)} = p^{(1)}P = p^{(0)}PP = p^{(0)}P^2$$
$$p^{(3)} = p^{(2)}P = p^{(0)}P^2P = p^{(0)}P^3 \quad \text{etc.}$$

Similarly, the nth step probability distribution (i.e., the distribution after the first n-steps) is denoted by

$$p^{(n)} = p^{(0)}P^n = \cdots p^{(n-1)}P$$

2.15 Stationary Distribution of Regular Markov Chains

Let P be a regular transition matrix of a Markov chain. Then in the long run, the probability that any state a_j occurs is approximately equal to the component v_j of the unique fixed probability vector v of P.
In other words, **Stationary distribution of a Markov chain is the unique fixed probability vector v** of the regular transition matrix P of the Markov chain because every sequence of probability distributions approaches v.

2.16 Absorbing States

A state a_i of a Markov chain is said to be an absorbing state if the system remains in the state a_i once it enters there, i.e., a state a_i is absorbing if $p_{ii} = 1$. Thus once a Markov chain enters such an absorbing state, it is destined there to remain forever. In other words the i th row in P has 1 at the main diagonal (i, i) position and zeros everywhere else.

Problem 60. *Prove that the Markov Chain whose t.p.m.* $P = \begin{bmatrix} 0 & \frac{2}{3} & \frac{1}{3} \\ \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 0 \end{bmatrix}$ *is irreducible. Find the corresponding stationary probability vector.*

We shall show that $\boldsymbol{P}$ is a regular stochastic matrix.
For convenience we shall write the given matrix in the form

$$\boldsymbol{P} = \frac{1}{6}\begin{bmatrix} 0 & 4 & 2 \\ 3 & 0 & 3 \\ 3 & 3 & 0 \end{bmatrix}$$

Consider $\boldsymbol{P^2} = \frac{1}{36}\begin{bmatrix} 0 & 4 & 2 \\ 3 & 0 & 3 \\ 3 & 3 & 0 \end{bmatrix}\begin{bmatrix} 0 & 4 & 2 \\ 3 & 0 & 3 \\ 3 & 3 & 0 \end{bmatrix} = \frac{1}{36}\begin{bmatrix} 18 & 6 & 12 \\ 9 & 21 & .6 \\ 9 & 12 & 15 \end{bmatrix}$

Since all the entries in $\boldsymbol{P^2}$ are positive we conclude that the t.p.m $\boldsymbol{P}$ is regular. Hence the Markov chain having t.p.m $\boldsymbol{P}$ is irreducible.

Next we shall find the fixed probability vector of $\boldsymbol{P}$. If $\boldsymbol{v} = (\boldsymbol{v_1}, \boldsymbol{v_2}, \boldsymbol{z})$ we shall find $\boldsymbol{v}$ such that $\boldsymbol{vP} = \boldsymbol{v}$ where $\boldsymbol{v_1} + \boldsymbol{v_2} + \boldsymbol{v_3} = \mathbf{1}$.

That is $[\boldsymbol{v_1}, \boldsymbol{v_2}, \boldsymbol{v_3}] \cdot \frac{1}{6}\begin{bmatrix} 0 & 4 & 2 \\ 3 & 0 & 3 \\ 3 & 3 & 0 \end{bmatrix} = [\boldsymbol{v_1}, \boldsymbol{v_2}, \boldsymbol{v_3}] \Rightarrow \quad \frac{1}{6}[\boldsymbol{3v_2} + \boldsymbol{3v_3}, \boldsymbol{4v_1} + \boldsymbol{3v_3}, \boldsymbol{2v_1} + \boldsymbol{3v_2}] = [\boldsymbol{v_1}, \boldsymbol{v_2}, \boldsymbol{v_3}] \Rightarrow \quad \boldsymbol{3v_2} + \boldsymbol{3v_3} = \boldsymbol{6v_1}; \boldsymbol{4v_1} + \boldsymbol{3v_3} = \boldsymbol{6v_2}; \boldsymbol{2v_1} + \boldsymbol{3v_2} = \boldsymbol{6v_3}$ Solving thesebyusing $\boldsymbol{v_1} + \boldsymbol{v_2} + \boldsymbol{v_3} = \mathbf{1}$ we obtain $\boldsymbol{v_1} = \mathbf{1/3}, \boldsymbol{v_2} = \mathbf{10/27}, \boldsymbol{v_3} = \mathbf{8/27}$

Thus $\boldsymbol{v} = \mathbf{(1/3, 10/27, 8/27)}$ is the required stationary probability vector.

Problem 61. *Every year, a man trades his car for a new car. If he has a Maruti, he trades it for an Ambassador. If he has an Ambassador, he trades it for a Santro. However, if he has a Santro, he is just as likely to trade it for a new Santro as to trade it for a Maruti or an Ambassador. In 2000 he bought his first car, which was a Santro.*
Find the probability that he has
(i) 2002 Santro
(ii) 2002 Maruti
(iii) 2003 Ambassador
(iv) 2003 Santro
(ii) In the long run, how often will he have a Santro.

Solution: (i) Define 3 states $\boldsymbol{a_1}, \boldsymbol{a_2}, \boldsymbol{a_3}$ as follows $\boldsymbol{a_1}$: state of having Maruti car(M), $\boldsymbol{a_2}$: having Ambassador(A), $\boldsymbol{a_3}$: having Santro(S).
Then the state space is $\boldsymbol{I} = \{\underset{(M)}{\boldsymbol{a_1}}, \underset{(A)}{\boldsymbol{a_2}}, \underset{(S)}{\boldsymbol{a_3}}\} = \{\boldsymbol{M}, \boldsymbol{A}, \boldsymbol{S}\}$

Then the transition probability matrix(tpm) is

$$P = \begin{array}{c} \\ a_1 \\ a_2 \\ a_3 \end{array} \begin{array}{c} \begin{array}{ccc} a_1 & a_2 & a_3 \end{array} \\ \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{pmatrix} \end{array}$$

In 2000 (Initial state): $p^{(0)} = (0, 0, 1) = \underset{(M)}{0}, \underset{(A)}{0}, \underset{(S)}{1}$ since he has Santro car in 2000 (his first purchase).

In 2001(after one step(year)), let us find $p^{(1)}$

$$p^{(1)} = p^{(0)}P = (0\,0\,1)\begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{pmatrix}$$

$$= \begin{pmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{pmatrix}$$

In 2002((after two steps), let us find $p^{(2)}$

$$p^{(2)} = p^{(1)}P = \begin{pmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{pmatrix}\begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{pmatrix}$$

$$= \underset{(M)}{\frac{1}{9}}\ \underset{(A)}{\frac{4}{9}}\ \underset{(S)}{\frac{4}{9}}$$

Hence (i) Probability that he has 2002 Santro = $\frac{4}{9}$

(ii) Probability that he has 2002 Maruthi = $\frac{1}{9}$

In 2003, let us find $p^{(3)}$

$$p^{(3)} = p^{(2)}P = \begin{pmatrix} \frac{1}{9} & \frac{4}{9} & \frac{4}{9} \end{pmatrix}\begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{pmatrix}$$

$$= \underset{(M)}{\frac{4}{27}}\ \underset{(A)}{\frac{7}{27}}\ \underset{(S)}{\frac{16}{27}}$$

Hence (iii) Probability that he has 2003 Ambassador = $\frac{7}{27}$

(iv) Probability that he has 2003 Santro = $\frac{16}{27}$

(v) To discover what happens in the long run, we must find a fixed probability vector

$v = (v_1,\ v_2,\ v_3)$ of P, such that

$$\left.\begin{aligned} vP &= v \\ \sum v_i &= 1 \end{aligned}\right\}$$

i.e.

$$\left.\begin{aligned} (v_1,\ v_2,\ v_3)\begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{pmatrix} &= (v_1,\ v_2,\ v_3) \\ v_1 + v_2 + v_3 &= 1 \end{aligned}\right\}$$

or

$$\left.\begin{aligned} v_1(0) + v_2(0) + \frac{1}{3}v_3 &= v_1 \\ v_1 + v_2(0) + \frac{1}{3}v_3 &= v_2 \\ v_1(0) + v_2 + \frac{1}{3}v_3 &= v_3 \\ v_1 + v_2 + v_3 &= 1 \end{aligned}\right\}$$

Rearranging, we get

$$-v_1 + (0)v_2 + \frac{1}{3}v_3 = 0 \qquad (1)$$

$$v_1 - v_2 + \frac{1}{3}v_3 = 0 \qquad (2)$$

$$(0)v_1 + v_2 - \frac{2}{3}v_3 = 0 \qquad (3)$$

$$v_1 + v_2 + v_3 = 1 \qquad (4)$$

When solving the system of equations, please note that all four equations (equations (1), (2), (3), and (4)) are essential for obtaining a meaningful solution. Equation (4) provides information required for a probability vector and ensures a unique and non-trivial solution. Hence select any two equations from (1), (2) and (3) and solve them together with (4), and then verify the answer using the equation, which is not used. For example, by solving (1), (2), and (4), we find the solution:

$$v_1 = \frac{1}{6}, v_2 = \frac{1}{3}, v_3 = \frac{3}{6} = \frac{1}{2}$$

To confirm the accuracy of this solution, substitute these values into equation (3) and check if it holds true.

Thus

$$v = (v_1,\ v_2,\ v_3) = \underset{(M)}{\frac{1}{6}}\ ,\ \underset{(A)}{\frac{1}{3}}\ ,\ \underset{(S)}{\frac{1}{2}}$$

In the long run, probability that will he have a Santro $=\frac{1}{2}$
i.e. He has a Santro 50% of the time.

Problem 62. *Three boys A, B, C are throwing ball to each other. A always throws the ball to B and B always throws the ball to C. C is just likely to throw the ball to B as to A. If C was the first person to throw the ball find probabilities that after three throws (i) A has the ball (ii) B has the ball (iii) C has the ball*

Solution : Here, the state space is $\boldsymbol{I} = \{\text{A has the ball, B has the ball, C has the ball}\}$ and the associated t.p.m. is as follows.

$$\boldsymbol{P} = \begin{array}{c} \\ (A) \\ (B) \\ (C) \end{array} \begin{array}{c} \begin{array}{ccc} (A) & (B) & (C) \end{array} \\ \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ \frac{1}{2} & \frac{1}{2} & 0 \end{bmatrix} \end{array}$$

Initially if $\boldsymbol{C}$ has the ball, the associated initial probability vector is given by

$$\boldsymbol{p^{(0)}} = \underset{(A)}{(0}, \underset{(B)}{0}, \underset{(C)}{1})$$

Since the probabilities are desired after three throws we have to find $\boldsymbol{p^{(3)} = p^{(0)} P^3}$ (or we can also find $\boldsymbol{p^{(1)} = p^{(0)} P}$, $\boldsymbol{p^{(2)} = p^{(1)} P}$ and then $\boldsymbol{p^{(3)} = p^{(2)} P}$)
By finding the powers of $\boldsymbol{P}$, we get

$$\boldsymbol{P^3} = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{2} \end{bmatrix}$$

$$\therefore \quad \boldsymbol{p^{(3)} = p^{(0)} P^3}$$

$$= (0 \ 0 \ 1) \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{2} \end{bmatrix}$$

$$= \begin{bmatrix} \underset{(A)}{\frac{1}{4}}, \underset{(B)}{\frac{1}{4}}, \underset{(C)}{\frac{1}{2}} \end{bmatrix}$$

Thus after three throws the probability that the ball is with $\boldsymbol{A}$ is $\frac{1}{4}$, with $\boldsymbol{B}$ is $\frac{1}{4}$ and with $\boldsymbol{C}$ is $\frac{1}{2}$.

Problem 63. *A gambler's luck follows a pattern such that if he wins a game, the probability of winning the next game is 0.6, and if he loses a game, the probability of losing the next game is 0.7. There is an even chance that the gambler wins the first game. What is the probability that he wins (i) the second game, (ii) the third game (iii) In the long run, how aften he will win?*

Solution : State space is : $I = \{$ Win (W), Lose $(L)\}$ and the associated transition probability matrix is as follows.

$$P = \begin{matrix} \\ (W) \\ (L) \end{matrix} \begin{bmatrix} (W) & (L) \\ 0.6 & 0.4 \\ 0.3 & 0.7 \end{bmatrix} = \frac{1}{10}\begin{bmatrix} 6 & 4 \\ 3 & 7 \end{bmatrix}$$

Given that Probability of winning the first game is $\frac{1}{2}$.
Hence **Initial State(first game)** probability vector is

$$p^{(0)} = (\frac{1}{2}, \frac{1}{2})$$

(i) Now for the **second game**,

$$p^{(1)} = p^{(0)}P = (\frac{1}{2}, \frac{1}{2})\begin{bmatrix} 0.6 & 0.4 \\ 0.3 & 0.7 \end{bmatrix}$$

$$= \frac{1}{2}[1, 1] \cdot \frac{1}{10}\begin{bmatrix} 6 & 4 \\ 3 & 7 \end{bmatrix}$$

$$= \frac{1}{20}[9, 11]$$

Hence

$$p^{(1)} = [\underset{(W)}{\frac{9}{20}}, \underset{(L)}{\frac{11}{20}}$$

Thus the probability of he winning the second game is $\frac{9}{20}$.
(ii) For the **third game**,

$$p^{(2)} = p^{(1)}P = \frac{1}{20}[9, 11] \cdot \frac{1}{10}\begin{bmatrix} 6 & 4 \\ 3 & 7 \end{bmatrix}$$

$$= \frac{1}{200}[87, 113]$$

Hence

$$p^{(2)} = [\underset{(W)}{\frac{87}{200}}, \underset{(L)}{\frac{113}{200}}]$$

Thus the probability of winning the third game is $\frac{87}{200}$. (iii) In the long run, we shall find the fixed probability vector

$$v = (v_1, v_2) \text{ such that } vP = v \text{ where } v_1 + v_2 = 1$$

That i.e. $[v_1, v_2]\frac{1}{10}\begin{bmatrix} 6 & 4 \\ 3 & 7 \end{bmatrix} = [v_1, v_2]$

and $v_1 + v_2 = 1$

$$\begin{aligned} 6v_1 + 3v_2 &= 10v_1 \\ \Rightarrow 4v_1 + 7v_2 &= 10v_2 \\ v_1 + v_2 &= 1 \end{aligned}$$

or

$$\begin{aligned} -4v_1 + 3v_2 &= 0 \quad (1) \\ 4v_1 - 3v_2 &= 0 \quad (2) \\ v_1 + v_2 &= 1 \quad (3) \end{aligned}$$

Select any one equation from (1) and (2), and solve it with (3).

For example, by solving (1) and (43), we find the solution:

$$v_1 = \frac{3}{7} \text{ and } v_2 = \frac{4}{7}$$

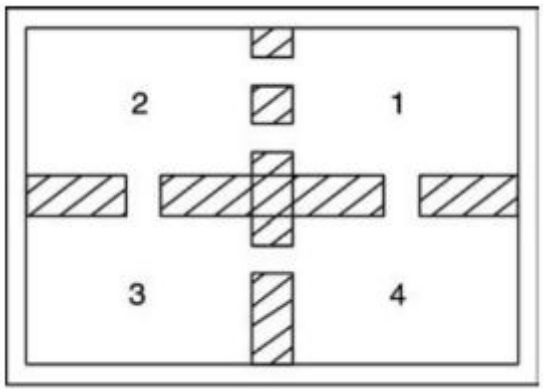
To confirm the accuracy of this solution, substitute these values into equation (2) and check if it holds true.

Hence

$$v = \underset{(W)\ \ (L)}{\left[\frac{3}{7}, \frac{4}{7}\right]}$$

Thus in the long run he wins $\frac{3}{7} = 42.857\%$ of the time.

Problem 64. *The following figure shows four compartments with door leading from one to another. A mouse in any compartment is equally likely to pass through each of the doors of the compartment. Find the transition matrix of the Markov chain. Draw the transition diagram.*

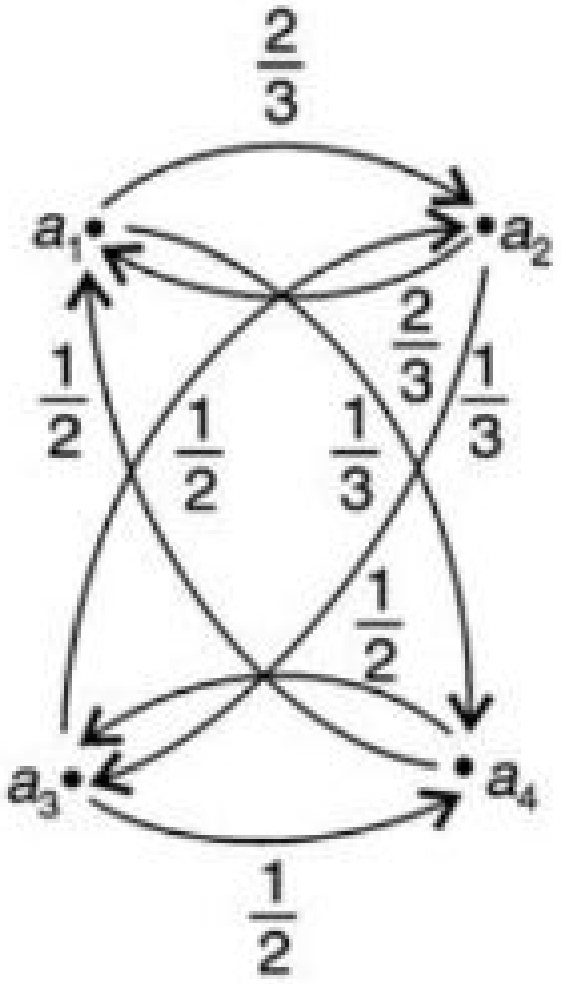


Solution : The 4 rooms are considered as four states say $1, 2, 3, 4$. Since mouse is moving, it does not stay in the same room. From room 1 it can go to 4 or 2 with

probability $\frac{1}{3}$ or $\frac{2}{3}$. It can not go from 1 to 3 . Then the first row consists of $0, \frac{2}{3}, 0, \frac{1}{3}$. Thus the transition matrix is

$$\begin{array}{c} \\ (1) \\ (2) \\ (3) \\ (4) \end{array} \begin{pmatrix} (1) & (2) & (3) & (4) \\ 0 & \frac{2}{3} & 0 & \frac{1}{3} \\ \frac{2}{3} & 0 & \frac{1}{3} & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 \end{pmatrix}$$

The transition diagram is



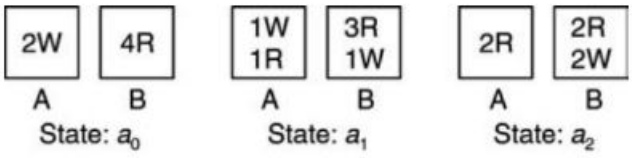
Problem 65. *Suppose an urn A contains 2 white marbles and urn B contains 4 red marbles. At each step of the process, a marble is selected at random from each urn and the two marbles selected are interchanged. Let X_n denote the number of red marbles in urn A after n interchanges.*

(i) Find the transition matrix P.

(ii) What is the probability that there are 2 red marbles in urn A after 3 steps.

(iii) In the long run, what is the probability that there are 2 red marbles in urn A.
(iv) What is the stationary distribution of the system.

Solution: Here, the state space is the no. of red marbles in urn A.
There are three states $I = 0, 1, 2$.
Since the number of marbles in the urn A is always 2 the possibilities can be represented by the following figure.



(i) **Transition matrix :** If the system is in the state a_0, then a white marble from A and a red from B must be selected and interchanged, so that the system will now move to state a_1. Accordingly the first row of the transition matrix (T.M.) is $(0, 1, 0)$.

Now suppose the system is in a_1. It can move to state a_0, iff red from A and white from B with probability $\frac{1}{2} \cdot \frac{1}{4} = \frac{1}{8}$. Thus $p_{10} = \frac{1}{8}$.

The system can move from a_1 to a_2, iff white from A and red from B with probability $\frac{1}{2} \cdot \frac{3}{4} = \frac{3}{8}$ i.e., $p_{12} = \frac{3}{8}$,

probability that system will remain in a_1 itself is $1 - \frac{1}{8} - \frac{3}{8} = \frac{1}{2}$.
(Note: white from A and white from B with probability $\frac{1}{2} \cdot \frac{1}{4} = \frac{1}{8}$ or red from A and red from B with probability $\frac{1}{2} \cdot \frac{3}{4} = \frac{3}{8}$. Thus the probability that system will remain in state a_1 itself is $\frac{1}{8} + \frac{3}{8} = \frac{1}{2}$).
Thus the 2nd row of T.M. is $\left(\frac{1}{8}, \frac{1}{2}, \frac{3}{8}\right)$.

Finally, suppose the system is in state a_2. Note that the system can never move from state a_2 to a_0. However, it may remain in a_2 itself, if a red from A and red from B is chosen. In this case the probability is $\frac{1}{1} \cdot \frac{2}{4} = \frac{1}{2}$.

Lastly, if a red from A and white from B is chosen, then system moves from a_2 to a_1 with probability $\frac{2}{4} = \frac{1}{2}$. Thus third row of the T.M. is $\left(0, \frac{1}{2}, \frac{1}{2}\right)$.

The Transition Matrix is Transition matrix P:

$$P = \begin{bmatrix} 0 & 1 & 0 \\ \frac{1}{8} & \frac{1}{2} & \frac{3}{8} \\ 0 & \frac{1}{2} & \frac{1}{2} \end{bmatrix}$$

(ii) The system starts in state a_0, so that $p^{(0)} = (1, 0, 0)$ is the initial state. Now

$$p^{(1)} = p^{(0)}P = \begin{pmatrix} 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} 0 & 1 & 0 \\ \frac{1}{8} & \frac{1}{2} & \frac{3}{8} \\ 0 & \frac{1}{2} & \frac{1}{2} \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 \end{pmatrix}$$

$$p^{(2)} = p^{(1)}P = \begin{pmatrix} 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} 0 & 1 & 0 \\ \frac{1}{8} & \frac{1}{2} & \frac{3}{8} \\ 0 & \frac{1}{2} & \frac{1}{2} \end{pmatrix} = \begin{pmatrix} \frac{1}{8} & \frac{1}{2} & \frac{3}{8} \end{pmatrix}$$

$$p^{(3)} = p^{(2)}P = \begin{pmatrix} \frac{1}{8} & \frac{1}{2} & \frac{3}{8} \end{pmatrix} \begin{pmatrix} 0 & 1 & 0 \\ \frac{1}{8} & \frac{1}{2} & \frac{3}{8} \\ 0 & \frac{1}{2} & \frac{1}{2} \end{pmatrix} = \begin{pmatrix} \frac{1}{16} & \frac{9}{16} & \frac{6}{16} \end{pmatrix}$$

Probability that there are two red marbles in A(i.e., in state a_2) after three steps is $\frac{6}{16} = \frac{3}{8}$. (iii) To study the system in the long run, we should find a unique fixed probability vector $v = (v_1, v_2, v_3)$ of the transition matrix P.

Let v be $(v_1, v_2, v_3) = (v_1, v_2, 1 - v_1 - v_2)$ ($\because v_1 + v_2 + v_3 = 1$). Then

$$vP = v$$

$$\begin{pmatrix} v_1 & v_2 & v_3 \end{pmatrix} \begin{pmatrix} 0 & 1 & 0 \\ \frac{1}{8} & \frac{1}{2} & \frac{3}{8} \\ 0 & \frac{1}{2} & \frac{1}{2} \end{pmatrix} = \begin{pmatrix} v_1 & v_2 & v_3 \end{pmatrix}$$

Solving

$$\frac{1}{8}v_2 = v_1 \text{ or } \quad v_2 = 8v_1$$

$$v_1 + \frac{1}{2}v_2 + \frac{1}{2}v_3 = v_2 \quad \text{or} \quad 2v_1 - v_2 + v_3 = 0$$

$$\frac{3}{8}v_2 + \frac{1}{2}v_3 = v_3 \quad \text{or} \quad 3v_2 = 4v_3$$

Now $3v_2 = 4v_3 = 4(1 - v_1 - v_2)$

$$7v_2 = 4 - 4v_1 \quad \text{or} \quad 56v_1 + 4v_1 = 4$$

$$\therefore \quad v_1 = \frac{4}{60}, v_2 = \frac{8}{15}, v_3 = \frac{6}{15}$$

Therefore, the fixed vector

$$t = \begin{pmatrix} \frac{1}{15} & \frac{8}{15} & \frac{6}{15} \end{pmatrix}$$

Hence the system in the long run stays in the state $a_2, 40\%$ of the time $\left(\frac{6}{15} = \frac{2}{5}\right)$ (i.e., there will be 2 red in $A, 40\%$ of the time). (iv) The fixed unique probability vector $v = \left(\frac{1}{15}, \frac{8}{15}, \frac{6}{15}\right)$ is the stationary distribution, since P^n approaches v, in the long run.

2.17 Question Bank

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

X / Y	-4	2	7
1	1/8	1/4	1/8
5	1/4	1/8	1/8

Determine (i) Marginal Distributions of X and Y (ii) Covariance of X and Y (iii) Correlation of X and Y [VTU Dec 2018, July 2017]

2) The joint distribution of two random variables X and Y is given below.

X / Y	-3	2	4
1	0.1	0.2	0.2
3	0.3	0.1	0.1

Determine (i) Marginal probability Distributions of X and Y (ii) Covariance of X and Y (iii) Correlation of X and Y (iv) Are X and Y independent? [VTU Jan 2014] Ans : $\mathrm{COV}(X, Y) = -1.2$, $\rho(X, Y) = -0.4$ (X and Y are not independent random variables)

3) The joint distribution of two random variables X and Y is given below.

X / Y	1	3	6
1	1/9	1/6	1/18
3	1/6	1/4	1/12
6	1/18	1/12	1/36

Determine the Marginal Distribution of X and Y.

Also, find whether X and Y are independent. [VTU Model 2020, July 2013]

4) X and Y are independent random variables. X take values, $2, 5, 7$ with probability $\frac{1}{2}, \frac{1}{4}$ and $\frac{1}{4}$, respectively. Y takes values $3, 4, 5$ with the probability $\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$

(a) Find the joint probability distribution of X and Y.

(b) Show that the covariance of X and Y is equal to zero.

(c) Find the probability distribution of Z = X + Y

5) The joint distribution of two random variables X and Y is given below.

X / Y	2	3	4
1	0.06	0.15	0.09
2	0.14	0.35	0.21

Determine the Marginal Distribution of X and Y.
Also, find whether X and Y are independent. [VTU Dec 2012]

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

X / Y	-2	-1	4	6
1	0.1	0.2	0	0.3
2	0.2	0.1	0.1	0

Determine (i) Marginal Distributions of X and Y (ii) Covariance of X and Y (iii) Correlation of X and Y [VTU Jan 2020,June 2019, Dec 2010]

7) A fair coin is tossed thrice. The random variables X and Y are defined as follows: $\boldsymbol{X = 0}$ 1 according as head or tail occurs on the first toss, y= no. of heads.
(i) Determine the marginal probability distribution of X and Y (ii)Determine the joint distribution of X and Y (iii) Determine E(X), E(Y) and E(XY) (iv) Determine $\boldsymbol{\sigma_x}$, $\boldsymbol{\sigma_y}$ [VTU]

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

X / Y	1	3	9
2	1/8	1/24	1/12
4	1/4	1/4	0
6	1/8	1/24	1/12

Determine (i) Marginal Distributions of X and Y (ii) Covariance of X and Y
[VTU Dec 2018, June 2011]

9) Compute (i) $\boldsymbol{P(X = 1, Y = 2)}$ (ii) $\boldsymbol{P(X \geq 1, Y \geq 2)}$ (iii) $\boldsymbol{P(X \leq 1, Y \leq 2)}$ (iv) $\boldsymbol{P(X + Y \geq 2)}$, using the folowing probability distribution of X and Y:

X/Y	0	1	2	3	Sum
0	0	$\frac{1}{8}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{2}$
1	$\frac{1}{8}$	$\frac{1}{4}$	$\frac{1}{8}$	0	$\frac{1}{2}$
Sum	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$	1

[VTU]

Ans : (i)$\frac{1}{8}$ (ii) $\frac{1}{2}$ (iii)$\frac{7}{8}$ (iv) $\frac{3}{4}$

10) A fair coin is tossed 4 times. Let X denote the no. of heads occuring and let Y denote the longest string of heads occurring. Find the joint distribution of X and Y. [VTU June 2011]

11) The joint distribution of two discrete random variables X and Y is $f(x, y) = k(2x + y)$ where x and y are integers such that $0 \leq x \leq 2, 0 \leq y \leq 3$. Find (i) The value of k (ii) Marginal distributions of X and Y (iii) Are X and Y independent ? [VTU Jan 2018]

12) X and Y are independent random variables. X takes the values $1, 2$ with probability $0.7, 0.3$ each and y takes the values $-2, 5, 8$ with probabilities $0.3, 0.5, 0.2$. Find the joint distribution of X and Y. Hence find $COV(X, Y)$. [VTU Jan 2018]

13) X and Y are independent random variables. X take values, $2, 5, 7$ with probability $\frac{1}{2}, \frac{1}{4}$ and $\frac{1}{4}$, respectively. Y takes values $3, 4, 5$ with the probability $\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$

(a) Find the joint probability distribution of X and Y.

(b) Show that the covariance of X and Y is equal to zero.

(c) Find the probability distribution of Z = X + Y

14) Determine (i) marginal distribution (ii) covariance between the discrete random variables X and Y, of the jont probsbility distribution

$\mathbb{X}\backslash$Y	3	4	5
2	1/6	1/6	1/6
5	1/12	1/12	1/12
7	1/12	1/12	1/12

(VTU Model 2020) Ans : (i) $p(x) = \frac{1}{2}, \frac{1}{4}, \frac{1}{4}$, $p(y) = \frac{1}{3}, \frac{1}{3}, \frac{1}{3}$ (ii) $COV(X, Y) = 0$

15) The random variable X takes values 0,1,2 with probability 0.3,0.3,0.4 and the random variable Y takes values 1,2,3 with probability $0.2, 0.2, 0.6$. If X and Y are independent random variables, (a) find the joint probability distribution of X and Y and (b) verify that $\text{Cov}(X, Y) = 0$

Ans :

X/Y	1	2	3
0	0.06	0.06	0.18
1	0.06	0.06	0.18
2	0.08	0.08	0.24

$\mu_X = 1.1, \mu_Y = 2.4, E[XY] = 2.64, Cov(X, Y) = 0$

16) Find the unique fixed probability vector for the regular stochastic matrix, $A = \begin{bmatrix} 0 & 1 & 0 \\ \frac{1}{6} & \frac{1}{2} & \frac{1}{3} \\ 0 & \frac{2}{3} & \frac{1}{3} \end{bmatrix}$

17) Find the unique fixed probability vector of (a) $A = \begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{3} & \frac{2}{3} & 0 \\ 0 & 1 & 0 \end{pmatrix}$,(b) $B = \begin{pmatrix} 0 & 1 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \end{pmatrix}$

Ans: (a) $\left(\frac{2}{9}, \frac{6}{9}, \frac{1}{9}\right)$
(b) $\left(\frac{5}{15}, \frac{6}{15}, \frac{4}{15}\right)$

18) Find the unique fixed probability vector for the regular stochastic matrix,
$A = \begin{bmatrix} 0 & \frac{3}{4} & \frac{1}{4} \\ \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & 1 & 0 \end{bmatrix}$ Ans: $v = \left(\frac{1}{4}, \frac{1}{2}, \frac{1}{2}\right)$

19) Define Stochastic Matrix. Find the unique fixed probability vector for the regular stochastic matrix, $A = \begin{bmatrix} \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 1 & 0 \end{bmatrix}$

20) Prove that the Markov Chain whose t.p.m. $P = \begin{bmatrix} 0 & \frac{2}{3} & \frac{1}{3} \\ \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 0 \end{bmatrix}$ is irreducible. Find the corresponding stationary probability vector.

21) Show that $P = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ \frac{1}{2} & \frac{1}{2} & 0 \end{bmatrix}$ is a regular stochastic matrix and find the corresponding unique fixed probability vector. Ans: In P^5 all the entries are positive. Hence P is a regular stochastic matrix.

$v = (1/5, 2/5, 2/5)$

22) Three boys A, B, C are throwing ball to each other. A always throws the ball to B and B always throws the ball to C. C is just likely to throw the ball to B as to A. If C was the first person to throw the ball find probabilities that after three throws (i) A has the ball (ii) B has the ball (iii) C has the ball Ans: Thus after three throws the probability that the ball is with A is $1/4$, with B is $1/4$ and with C is $1/2$.

23) A gambler's luck follows a pattern such that if he wins a game, the probability of winning the next game is 0.6, and if he loses a game, the probability of losing the next game is 0.7. There is an even chance that the gambler wins the first game. What is the probability that he wins (i) the second game, (ii) the third game (iii) In the long run, how often he will win? Ans : 9/20 and 87/200, $v = [3/7, 4/7]$

24) A student's study habits are as follows: If he studies one night, he is 70% sure not to study the next night. On the other hand, if he does not study one night, he is 60% sure not to study the next night as well. In the long run, how often does he study? Ans: v=$(4/11 , 7/11)$)

25) A software engineer goes to his work place every day by motor bike or by car. He never goes by bike on two consecutive days, but it he goes by car on a day then he is equally likely to go by car or by bike the next day. Find the t.p.m. of the Markov chain. If car is used on the first day of the week, find the probability that (i) bike is used (ii) cars is used, on the fifth day. Ans: $\frac{5}{16}, \frac{7}{16}$

26) A company executive changes his car every year. If he has a car of make A, he changes over to make B. From make B, he changes over to make C. However, if he has a car C, he is just as likely to change it for car of make A as to change it for a car of make B. If he had a car of make C in the year 2008,

(i) Find the probability that he had a car of (a) make A in 2010 (b) make C in 2010 (c) make C in 2011 (d) make B in 2011

(ii) In the long run, how often will he have a car of make C ?

Ans: (i) (a) Probability that he had 2002 Ford is $p_3^{(2)} = \frac{4}{9}$. (b) Probability that he had 2002 Tata is $p_1^{(2)} = \frac{1}{9}$. (c) Probability that he had 2003 Maruti is $p_2^{(3)} = \frac{7}{27}$. (d) Probability that he had 2003 Ford is $p_3^{(3)} = \frac{16}{27}$. (ii) $\mathbf{v} = \left(\frac{1}{6}, \frac{2}{6}, \frac{3}{6}\right)$. In the long run, the proportion of time he will have a Ford is $\frac{3}{6} = \frac{1}{2}$; i.e., 50% of the time.

27) Every year, a man trades his car for a new car. If he has a Maruti, he trades it for an. Ambassador. If he has an Ambassador, he trades it for a Santro. However, if he has a Santro, he is just as likely to trade if for a new Santro as to trade if for Maruti or an Ambassador. In 2000 he bought his first car which was Santro. Find the probability that he has (i) 2002 Santro (ii) 2002 Maruti

28) A man's smoking habits are as follows. If he smokes filter cigarettes one week, he switches to nonfilter cigarettes the next week with probability 0.2. On the other hand, if he smokes nonfilter cigarettes one week, there is a probability 0.7 that he switches to filter in the following week. In the long run how often does he smoke filter cigarettes? [VTU Jan 2018]

Ans: In the long run, he will smoke filter cigarettes $3/5$ or 60% of the time.

29) A saleman's territory consists of 3 cities A, B and C. He never sells in the same city on successive days. If he sells in city A, then the next day he sells in city B. However if he sells in either B or C, then the next day he is twice as likely to sell in city A as in other city. In the long run, how often does he sell in each of the cities. Ans: $v = \left(\frac{2}{5}, \frac{9}{20}, \frac{3}{20}\right)$. In the long run he sells 40% of time in city A, 45% in B, 15% of time in C.

30) There are 2 white marbles in box $\boldsymbol{A}$ and 3 red marbles in box $\boldsymbol{B}$. At each step of the process a marble is selected from each box and the two marbles selected are interchanged. Let the state $\boldsymbol{a_i}$ of the system be the number $\boldsymbol{i}$ of red marbles in box $\boldsymbol{A}$. (a) Find the transition matrix $\boldsymbol{P}$. (b) What is the probability that there are 2 red marbles in box $\boldsymbol{A}$ after 3 steps (c) In the long run, what is the probability that there are 2 red marbles in box $\boldsymbol{A}$. Ans : (b) Probability that there are 2 red marbles in box $\boldsymbol{A}$ after 3 steps is $\frac{5}{18}$. c) Fixed probability vector: $\boldsymbol{v = (0.1, 0.6, 0.3)}$. In the long run, $\boldsymbol{30\%}$ of the time, there will be 2 red marbles in box $\boldsymbol{A}$.

31) A habitual gambler is a member of two clubs $\boldsymbol{A}$ and $\boldsymbol{B}$. He visits either of the clubs everyday for playing cards. He never visits club $\boldsymbol{A}$ on two consecutive days. But, if he visits club $\boldsymbol{B}$ on a particular day, then the next day he is as likely to visit club $\boldsymbol{B}$ or club $\boldsymbol{A}$. Find the transition matrix of this Markov chain. Also, (a) show that the matrix is a regular stochastic matrix and find the unique fixed probability vector. (b) if the person had visited club B on Monday, find the probability that he visits club $\boldsymbol{A}$ on Thursday. Ans: $\boldsymbol{v} = (\frac{1}{3}, \frac{2}{3}), \frac{3}{8}$

32) Explain

(i) Regular and irregular Markov Chain

(ii) State Distribution and Higher Transition Probabilities

Module 3

Statistical Inference 1

3.1 Sampling

- Statistical Inference is a branch of Statistics which uses probability concepts to deal with uncertainty in decision making. There are a number of situations where in we come across problems involving decision making.
- It is necessary to draw some valid and reasonable conclusions concerning a large mass of individuals or things. Every individual or the entire group is known as population. Small part of this population is known as a sample.
- For example, consider the problem of buying 1 kilogram of rice, when we visit the shop, we do not check each and every rice grains stored in a gunny bag; rather we put our hand inside the bag and collect a sample of rice grains. Then analysis takes place. Based on this, we decide to buy or not. Thus, the problem involves studying whole rice stored in a bag using only a sample of rice grains.
- A large collection of individuals or numerical is regarded as population or universe.
- A finite subset of the universe is called a sample. The number of individuals in a sample is called a Sample Size (n). If $\boldsymbol{n \leq 30}$, then it is said to be small sample. Otherwise it is large sample.
- The selection of individual or item from the population in such a way that each has the same chance of being selected is called as random sampling.

- The statistical constant of the population (such as mean and standard deviation etc.) is referred as Parameter and the statistical constant of the Sample is referred as Statistic.

- For every sample of size n, we can compute quantities like mean, median, standard deviation etc., obviously these will not be the same. Suppose we group these characteristics according to their frequencies, the frequency distributions so generated are called Sampling Distributions. The sampling distribution of large samples is assumed to be a normal distribution.

- The standard deviation of a sampling distribution is also called as the standard error (SE).

- A statistical hypothesis is some assumption or statement about a population parameter, which may or may not be true, which we want to test on the basis of the evidence from a random sample.

3.2 Testing of Hypothesis:

- The procedure involves the following:
  1) **Null Hypothesis :** First we set up a definite statement about the population parameter which we call it as null hypothesis, denoted by $\boldsymbol{H_0}$. Null Hypothesis is the statement which is tested for possible rejection under the assumption that it is true.
  2) Next we set up another hypothesis called **alternate hypothesis** which is just complimentary to null hypothesis; denoted by $\boldsymbol{H_1}$. (i.e. The Null and Alternative hypothesis are complementary and the two together exhaust all possibilities regarding the value that the hypothesised parameters can assume).

3) Specify the level of significance: The level of significance is defined as the probability of rejecting Null hypothesis when it is true. The higher the level of significance, the higher the probability of rejecting a Null hypothesis when it is true.
Commonly used levels of significance in practice are $\mathbf{5\%(= 0.05)}$ and $\mathbf{1\%(= 0.01)}$.
4) **Test of Significance :** It enable us to decide on the basis of the sample results if the deviation between the observed sample statistic and the hypothetical parameter

value is significant. Use a test statistic to Calculate the value of the statistic from the given data.
5) Accept or Reject the Null Hypothesis. If the calculated value of the statistic is less than tabulated value, we accept $\boldsymbol{H_0}$. Otherwise we reject $\boldsymbol{H_0}$ and accept $\boldsymbol{H_1}$.

3.3 Testing of Hypothesis:

- The procedure involves the following:
  1) **Null Hypothesis :** First we set up a definite statement about the population parameter which we call it as null hypothesis, denoted by $\boldsymbol{H_0}$. Null Hypothesis is the statement which is tested for possible rejection under the assumption that it is true.
  2) Next we set up another hypothesis called **alternate hypothesis** which is just complimentary to null hypothesis; denoted by $\boldsymbol{H_1}$. (i.e. The Null and Alternative hypothesis are complementary and the two together exhaust all possibilities regarding the value that the hypothesised parameters can assume).
  3) Specify the level of significance: The level of significance is defined as the probability of rejecting Null hypothesis when it is true. The higher the level of significance, the higher the probability of rejecting a Null hypothesis when it is true.
  Commonly used levels of significance in practice are $\boldsymbol{5\%(=0.05)}$ and $\boldsymbol{1\%(=0.01)}$.
  4) **Test of Significance :** It enable us to decide on the basis of the sample results if the deviation between the observed sample statistic and the hypothetical parameter value is significant. Use a test statistic to Calculate the value of the statistic from the given data.
  5) Accept or Reject the Null Hypothesis. If the calculated value of the statistic is less than tabulated value, we accept $\boldsymbol{H_0}$. Otherwise we reject $\boldsymbol{H_0}$ and accept $\boldsymbol{H_1}$.

Example: Hypotheses in a Pharmaceutical Study Consider a scenario where a pharmaceutical company develops a new drug and claims that it reduces blood pressure in patients. The company may formulate the following hypotheses:

- **Null Hypothesis (H_0):** The new drug has no effect on reducing blood pressure. The average blood pressure for patients taking the new drug is the same as the average blood pressure for the general population.
- **Alternative Hypothesis (H_1 or H_a):** The new drug is effective in reducing blood pressure. The average blood pressure for patients taking the new drug is significantly lower than the average blood pressure for the general population.

In this example:

- The null hypothesis (H_0) represents the default assumption, suggesting no effect or no difference. - The alternative hypothesis (H_1 or H_a) represents the claim or the effect that the researcher is trying to provide evidence for.

3.4 Type I and Type II Errors:

When we test hypothesis, there are four possible outcomes.
(i) Null hypothesis is rejected when it is false.
(ii) Null hypothesis is rejected when it is true.
(iii) Null hypothesis accepted when it is false.
(iv) Null hypothesis is accepted when it is true.
These situations lead to the two types of errors and same is tabulated as follows:

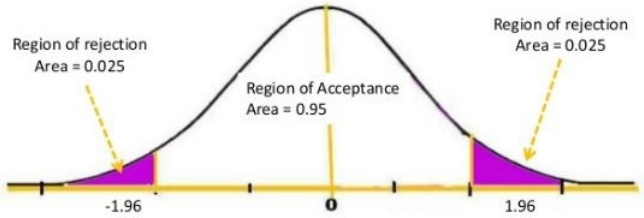
		Hypothesis	
		TRUE	FALSE
Decision	Accept	Right Decision	Type-II Error
	Reject	Type-I Error	Right Decision

A type I error (false-positive) occurs if an investigator rejects a null hypothesis that is actually true in the population; a type II error (false-negative) occurs if the investigator fails to reject a null hypothesis that is actually false in the population.

3.5 Critical Region:

The Critical Value serves as the basis for either Accepting or Rejecting a Hypothesis. When $\alpha = 0.05$, the Region of Rejection is 0.05 and the Region of Acceptance is 0.95

The value of Z corresponding of 5% level of significance is ± 1.96 and corresponding to 1% level of significance value of Z is ± 2.58. The set of Z-scores outside the range ± 1.96 and ± 2.58 constitutes the critical region of the hypothesis or the region of rejection or the region of significance at 5% and 1% level of significance respectively.



3.6 Confidence Interval:

A **confidence interval** is a statistical range that provides an estimate of the likely range of values for an unknown population parameter, such as the mean or proportion. It is calculated from a given set of sample data and is used to quantify the uncertainty or margin of error associated with the estimate.

The confidence interval is expressed as two values, an upper limit, and a lower limit, which form a range. The level of confidence represents the probability that the true parameter falls within the interval. Commonly used confidence levels are 95%, and 99%. For example, a 95% confidence interval implies that if we were to draw multiple samples and calculate a confidence interval from each, we would expect the true parameter to be within the interval for approximately 95% of those samples.

Let us suppose that we have a normal population with mean μ and S.D σ. If $\bar{x}$ is the sample mean of a random sample of size n the quantity z defined by

$$z = \frac{\bar{x} - \mu}{(\sigma/\sqrt{n})}$$

is called the Standard Normal Variate (S.N.V).

From the table of normal areas we find that 95% of the area lies between $z =$

-1.96 and $z = +1.96$. In other words we can say with 95% confidence that z liesbetween -1.96 and +1.96 . Further 5% level of significance is denoted by $z_{0.05}$. Thus we can write the verbal statement in the mathematical form,

$$-1.96 \leq \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \leq 1.96$$

$$\text{i.e., } \frac{-\sigma}{\sqrt{n}}(1.96) \leq \bar{x} - \mu \leq \frac{\sigma}{\sqrt{n}}(1.96)$$

$$\Rightarrow \mu \leq \bar{x} + \frac{\sigma}{\sqrt{n}}(1.96) \text{ and } \bar{x} - \frac{\sigma}{\sqrt{n}}(1.96) \leq \mu$$

Thus we can write by combining the two results in the form,

$$\bar{x} - 1.96\left(\frac{\sigma}{\sqrt{n}}\right) \leq \mu \leq \bar{x} + 1.96\left(\frac{\sigma}{\sqrt{n}}\right)$$

Confidence limits are the numbers at the upper and lower end of a confidence interval.

3.7 Tests for large samples:

z-test is a statistical test that is used to determine whether the mean of a sample is significantly different from a known population mean when the population standard deviation is known. It is particularly useful when the sample size is large (¿30).
Suppose that we have a normal population with mean μ and S.D. as σ. If $\bar{x}$ is the sample mean of a random sample of size n(where $n > 30$).
Some commonly used notations in sampling distributions are given below:

	Population	Sample
Size	N	n
Mean	μ	$\bar{x}$
Variance	σ^2	s^2
Standard Deviation	σ	s
Proportion	P	p

Suppose we take various samples each of size n from a population. If p and q are the probabilities of success and failure of each member of the sample, then the binomial distribution given by $(p+q)^n$ provides the sampling distribution of the number of successes in the sample with mean $\mu = np$ and standard deviation $\sigma = npq$.

- Mean (expected value) of number of successes $\mu = np$
- Standard deviation $\sigma = \sqrt{npq}$.

- Probable occurrence range(confidence interval) at 99% confidence level i.e. 1% significance level is given by: $np \pm 2.58\sqrt{npq}$
- Probable occurrence range(confidence interval) at 95% confidence level i.e. 5% significance level is given by: $np \pm 1.96\sqrt{npq}$

In case of proportion of successes, mean and standard deviation of proportion of successes are obtained by dividing each statistic by n.

- Mean (expected value) of proportion of successes $= \frac{np}{n} = p$
- Standard deviation of proportion $= \frac{\sqrt{npq}}{n} = \sqrt{\frac{pq}{n}}$
- Probable occurrence range(confidence interval) of the proportion at 99% confidence level i.e. 1% significance level is given by: $p \pm 2.58\sqrt{\frac{pq}{n}}$
- Probable occurrence range(confidence interval) of the proportion at 95% confidence level i.e. 5% significance level is given by: $p \pm 1.96\sqrt{\frac{pq}{n}}$

3.8 Test of significance of single mean

Suppose that we have a normal population with mean μ and S.D. as σ. When conducting a test of significance for a single mean, the aim is to determine whether the mean of a sample, $\bar{x}$ is significantly different from a known or hypothesized population mean, μ.

The test involves stating two hypotheses:

- **Null Hypothesis (H_0):** Assumes that there is no significant difference, and any observed difference is due to random chance.
- **Alternative Hypothesis (H_1 or H_a):** Asserts that there is a significant difference, suggesting that the observed result is not due to random chance alone.

The test statistic is a numerical value calculated from the sample data.
If $\bar{x}$ is the sample mean of a random sample of size n(where $n > 30$), then to test

whether the difference between sample mean and population mean is significant, the test statistic is :

$$\boxed{z = \frac{(\bar{x} - \mu)}{e}}$$

where standard Error, e is given by

$$e^2 = \frac{\sigma^2}{n}$$

Critical Values of z at 5% and 1% level of significance are respectively given by 1.96 and 2.58

If $|z| < 1.96$, we accept the hypothesis that there is no significant difference between the population mean and sample mean at 5% level of significance.

3.9 Test for the Mean Number of Successes in Normal Approximation to Binomial Distribution

Let x be the number of successes in n trials with probability p of success in each trial, $\mu = np$ is the expected number of successes and standard deviation is $\sigma = \sqrt{npq}$ (where $q = 1 - p$) and the test statistic is:

$$\boxed{z = \frac{x - \mu}{\sigma} = \frac{x - np}{e}}$$

where the standard Error, e is given by using the formula

$$e^2 = npq$$

This can be particularly useful in situations where a binomial distribution with a large sample size is involved, and normal approximation simplifies the analysis.

Therefore we have the following test of significance:

(i) If $|z| < 1.96$, difference between the observed and expected number of successes is not significant (at 5% level of significance).

(ii) If $|z| < 2.58$, difference between the observed and expected number of successes is not significant(at 1% level of significance).

(ii) If $|z| > 1.96$, difference is significant (at 5% level of significance).

(iii) If $|z| > 2.58$, difference is significant (at 1% level of significance).

3.10 Test of significance of single proportion

To test the significant difference between the sample proportion $\mathbf{p}$ and the Population proportion $\mathbf{P}$ we use the statistic (z-test) which is calculated as :

$$z = \frac{p - P}{e}$$

where the standard Error, e is given by using the formula

$$e^2 = \frac{PQ}{n}$$

and

- p is the sample proportion,
- P is the specified population proportion, and $P + Q = 1 \Rightarrow Q = 1 - P$
- n is the sample size.

The formulated Null and Alternative hypotheses are, $H_0 : P =$ a specified value and $H_1 : P \neq$ a specified value
After calculating the test statistic, the decision to accept the null hypothesis (H_0) depends on whether the test statistic falls within the critical region. This critical region is determined by comparing the calculated test statistic to tabulated values associated with the chosen level of significance.

3.11 Test of Significance for the Difference Between Means

When comparing the means of two independent samples from **two different populations**with large sizes, the z-test for the difference between means can be used.
To test whether the means are equal, We define the null and alternate hypothesis as $H_0 : \mu_1 = \mu_2$ and H_1: $\mu_1 \neq \mu_2$
and the test statistic is calculated as :

$$z = \frac{(\bar{x}_1 - \bar{x}_2)}{e}$$

with the standard error e calculated as:

$$e^2 = \left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

and

- $\bar{x}_1$ and $\bar{x}_2$ are the sample means,

- μ_1 and μ_2 are the population means,
- σ_1 and σ_2 are the population standard deviations, and
- n_1 and n_2 are the sample sizes.

The decision is made by comparing the z-statistic to critical values.
If $|z| < z_{0.05} = 1.96$, we accept the hypothesis that there is no significant difference between the population means μ_1 and μ_2 at 5% level of significance.
If $|z| < z_{0.01} = 2.58$, we accept the hypothesis that there is no significant difference between the population means μ_1 and μ_2 at 1% level of significance.

When comparing the means of two independent samples from **same population** with S.D. σ, the test statistic and the procedure are similar to the case of the different population. Here, $\sigma_1 = \sigma_2 = \sigma$. Hence to test whether the difference between the sample means $\bar{x}_1$ and $\bar{x}_2$ is significant or is merely due to fluctuations of sampling, the test statistic is calculated as :

$$z = \frac{\bar{x}_1 - \bar{x}_2}{e}$$

with the standard error e calculated as:

$$e^2 = \left(\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}\right)$$

or

$$e = \sigma\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

The decision is made by comparing the z-statistic to critical values.

3.12 Test of significance of Difference between two sample proportions

Given two large samples of sizes n_1, n_2 are taken from **two similar populations** giving sample proportions as p_1, p_2 respectively. To test whether the proportions P_1 and P_2 of the two populations are equal, We define the null and alternate hypothesis as
$H_0 : P_1 = P_2$ and H_1: $P_1 \neq P_2$
and the then we use the test statistic

$$z = \frac{p_1 - p_2}{e}$$

where the standard error e, of the difference between p_1 and p_2, is given by using

$$e^2 = PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$$

where (P) is the pooled sample proportion calculated by combining both samples and is given by

$$P = \frac{n_1p_1 + n_2p_2}{n_1 + n_2}, \quad Q = 1 - P$$

For the case of samples taken from the **two different type of populations**, the standard error e is given by using

$$e^2 = \left(\frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}\right)$$

where

- p_1 and p_2 are the sample proportions,
- Here, the null hypothesis (H_0) and alternative hypothesis $(H_1$ or $H_a)$ are stated as follows:

  $$H_0 : p_1 = p_2$$
  $$H_1 : p_1 \neq p_2$$

  where p_1 and p_2 are the sample proportions.

Problem 66. *A sample of 100 tyres is taken from a lot. The mean life of a tyre is found to be 39350 kms with a SD of 3260. Can it be considered as the true random sample from population with mean life of 40000 kms? (use 5% significance level). [VTU: DEC/JAN 16]*

Solution: First we shall set up null hypothesis, $H_0 : \mu = 40,000$,
Alternate hypothesis as $H_1 : \mu \neq 40,000$.
We consider that the problem follows a two tailed test and chose $\alpha = 5\%$. Then corresponding to this, tabulated value is 1.96.
Consider the expression for finding test criterion,
$z = \frac{\overline{x}-\mu}{\frac{\sigma}{\sqrt{n}}}$ (*) Here, $\mu = 40,000$, $\overline{X} = 39,350$ and $\sigma = 3,260, n = 100$.
S.E. $= \frac{\sigma}{\sqrt{n}} = \frac{3,260}{\sqrt{100}} = 326$.
Thus, from (8), $z = 1.994$.
As this value is slightly greater than 1.96, we reject the null hypothesis and conclude that sample has not come from a population of 40,000 kilometers.

Problem 67. *A die is tossed 960 times and 5 appear 184 times, is the die biased? [VTU:JUNE/JULY-15, 2006]*

Solution: Suppose the die is unbiased, given $n = 960$
Probability that 5 appears with one die $p = \frac{1}{6}$
$\therefore q = 1 - p = 1 - \frac{1}{6} = \frac{5}{6}$
Expected number of successes $np = \frac{1}{6} \times 960 = 160$
Observed value of successes $x = 184$ Standard deviation of simple sampling is

$$\sqrt{npq} = \sqrt{960 \times \frac{1}{6} \times \frac{5}{6}} = \sqrt{133.33} = 11.55$$

Hence $z = \frac{x-np}{\sqrt{npq}} = \frac{24}{11.55} = 2.078 > 1.96$
We reject the hypothesis at 5% level of significance and we conclude that the die is biased.

Problem 68. *In 324 throws of a six faced 'die' an odd number turned up 181 times. Is it reasonable to think that the die is an unbiased one at 1% level of significance.? [VTU July 2017]*

Solution: Lel us take H_0 : die is unbiased
and the probability of getting odd number is $p = \frac{3}{6} = \frac{1}{2} = 0.5$
since $p + q = 1, q = 1 - p = 0.5$
Expected number of successes $= np = 324 \times 0.5 = 162,$
$npq = (324)(0.5)(0.5) = 81$

$$Z = \frac{x - np}{\sqrt{npq}} = \frac{181 - 162}{\sqrt{81}} = \frac{19}{\sqrt{81}} = \frac{19}{9} = 2.11 < z_{0.01} = 2.58$$

Thus we can say that the die is unbiased.

Problem 69. *A coin is tossed 1000 times and head turns up 540 times. Test the hypothesis that the coin is an unbiased one. [VTU Model 2023, Dec 2018, Jan 2014]*

Solution: Let H_0 : The coin is unbiased.
and let $p =$ the probability of gettlng a head in one toss $= \frac{1}{2} = 0.5$
since $p + q = 1,\ q = 1 - p = 0.5$
Expected number of heads in 1000 tosses $= np = 1000 \times 0.5 = 500,$
$npq = (1000)(0.5)(0.5) = 250$
$\therefore$

$$z = \frac{x - np}{\sqrt{npq}} = \frac{540 - 500}{\sqrt{250}}$$

$$\Rightarrow z = \frac{40}{\sqrt{250}} = 2.53 < 2.58$$

Thus we can say that the coin is unbiased.

Problem 70. *A manufacturing company claims that at least* 95% *of its products supplied confirm to the specifications out of a sample of 200 products, 18 are defective. Test the claim at* 5% *Los.*

Solution: Set the null hypothesis $H_0; P = 95\% = 0.95$
Set the Alternative hypothesis $H_1 : P \neq 0.95$
The level of significance $\alpha = 0.05(5\%)$
$\therefore$ The test statistic is
$Z = \frac{p-P}{\sqrt{\frac{PQ}{n}}}$,
where $\mathrm{P} + \mathrm{Q} = 1 \Rightarrow \mathrm{Q} = 1 - \mathrm{P}$
Given that, 18 products are defective out of 200 sample products
$\therefore$ The total defective less products $= 200 - 18 = 182$

$$\therefore p = \frac{182}{200} = 0.91$$

$$\therefore z = \frac{0.91 - 0.95}{\sqrt{\frac{0.95 \times 0.05}{200}}}$$

$$\Rightarrow Z = -\frac{0.04}{\sqrt{0.0002375}} = -2.5955$$

At 5% level, the tabulated value of Z_α is 1.96 Since $|Z| = 2.5955 > 1.96$
Hence, the null hypothesis is rejected at 5% level of significance

Problem 71. *A stenographer claims that she can type at the rate of 120 words per minute. Can we reject her claim on the basis of 100 trails in which she demonstrates a mean of 116 words with a standard deviation of 15 words? Use 5% level of significance.* *(VTU Model 2020)*

Solution: The hypothesis to be tested is that $H_0 : \mu = 120$
$H_1 : \mu \neq 120$
Given $\bar{x} = 116, s = 15, n = 100$

$$Z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}, \text{ we get}$$

$$\Rightarrow Z = \left|\frac{116-120}{15/\sqrt{100}}\right|$$

$$\Rightarrow Z = 2.67 > z_{0.05} = 1.96$$

Hence H_0 is rejected and H_1 is accepted. i.e. $\mu \neq 120$

Problem 72. *In a large city A, 20% of a random sample of 900 school boys had a slight physical defect. In another large city B, 18.5% of a random sample of 1600*

school boys had the same defect. Is the difference between the proportions significant at 5% significance level? *(VTU Model 2023)*

Solution : Set the null hypothesis $\mathbf{H_0 : P_1 = P_2}$ Set the alternative hypothesis $\mathbf{H_1 : P_1 \neq P_2}$
Level of significance $\alpha = 0.05(5\%)$
The test statistic is

$$z = \frac{p_1 - p_2}{\sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

where $P + Q = 1 \Rightarrow Q = 1 - P$
Given $n_1 = 900, n_2 = 1600, p_1 = \frac{20}{100} = 0.2, p_2 = \frac{18.5}{100} = 0.185$

$$P = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{(900)0.2 + (1600)0.185}{900 + 1600} = 0.1904$$

$$\Rightarrow Q = 1 - P = 1 - 0.1904 = 0.8096$$

$$\Rightarrow Z = \frac{p_1 - p_2}{\sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{0.2 - 0.185}{\sqrt{(0.1904)(0.8096)\left(\frac{1}{900} + \frac{1}{1600}\right)}} = 0.9305$$

Critical value: At 5% level, the tabulated value of Z_α is 1.96 .
Conclusion: Since $|Z| = 0.9305 < 1.96$
Hence Null Hypothesis H_0 is accepted at 5% level of significance.
Hence there is no significant difference.

Problem 73. *In two large populations there are* 30% *and* 25% *respectively of fair haired people. Is this difference likely to be hidden in samples of 1200 and 900 respectively from the two populations?*

Solution : Here $p_1 = 0.3, p_2 = 0.25$ so that $p_1 - p_2 = 0.05$. $\therefore$

$$e^2 = \frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2} = \frac{0.3 \times 0.7}{1200} + \frac{0.25 \times 0.75}{900}$$

so that

$$e = 0.0195$$

$\therefore$

$$z = \frac{p_1 - p_2}{e} = \frac{0.05}{0.0195} = 2.5 \text{ nearly}$$

Hence it is unlikely that the real difference will be hidden.

Problem 74. *A sample of 100 students is taken from a large population. The mean height of the students in this sample is 160cm. Can it be reasonably regarded that*

this sample is from a population of mean 165 cm and S.D 10 cm? (VTU Model 2020)

Solution: Given that $n = 100, \quad \bar{x} = 160, \quad \sigma = 10$ and $\mu = 165$
Null Hypothesis: $H_0 : \mu = 165$ i.e., there is no difference between sample mean and population mean.
Alternative Hypothesis: $H_1 : \mu \neq 165$ (Two tailed alternative)
The test statistic is given by

$$z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} = \frac{160-165}{\frac{10}{\sqrt{100}}} = -5$$

$\therefore \quad z = -5$ [Calculated value]
At 5% significance level the tabulated value for z_α is 1.96
But $|z| > 1.96$.
so we reject H_0
Conclusion:
There is a significant difference between the sample mean and population means.

Problem 75. *The mean life time of a sample of 100 fluorescent tube lights manufactured by a company is found to be 1570 hrs with a standard deviation of 120 hrs. Test the hypothesis that the mean life-time of the lights produced by the company is 1600 hrs at 0.01 level of significance. (VTU Model 2020)*

Solution: Let H_0 :the mean life-time of the lights produced by the company is 1600
Given $\bar{x} = 1570, n = 100, s = 120, \mu = 1600$
Apply the formula to calculate z score:

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{1570 - 1600}{120/\sqrt{100}} = -2.5$$

$|z| = 2.5 < z_{0.01} = 2.58$ Therefore, we accept the hypothesis H_0 at 0.01 level of significance.

Problem 76. *It is claimed that a random sample of 49 tyres has a mean life of* $15,200\text{kms}$. *Is the sample drawn from a population whose mean is* $15,150\text{kms}$ *and whose standard deviation is 1,200 kms? Test the significance at 0.05 level.(VTU Model 2020)*

Solution: Given $n = 49, \ \bar{x} = 15200, \ \mu = 15150$ and $\sigma = 1200$
Null hypothesis $H_0 : \mu = 15200$
Alternate hypothesis $H_1 : \mu \neq 15200$

Level of significance is $\alpha : 0.05$
The test statistic is $Z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$

$$\Rightarrow Z = \frac{15200-15150}{1200/\sqrt{49}}$$
$$\Rightarrow Z = 0.2917$$

since $-1.96 < Z < 1.96$,, (ie. $|z| < Z_{0.05} = 1.96$), we accept the null hypothesis.

Problem 77. *In a sample of 600 men from a certain city, 450 are found smokers. In another sample of 900 men from another city, 450 are smokers. Do the data indicate that the cities are significantly different with respect to the habit of smoking among men. Test at 5% significance level.* *(VTU Model 2023)*

Solution : For sample 1, we have that the sample proportion is $\hat{p}_1 = \frac{X_1}{N_1} = \frac{450}{600} = 0.75$.

For sample 1, we have that the sample proportion is $\hat{p}_2 = \frac{X_2}{N_2} = \frac{450}{900} = 0.5$.

The value of the pooled proportion is computed as

$$\frac{X_1 + X_2}{N_1 + N_2} = \frac{400 + 450}{600 + 900} = 0.56$$

The following null and alternative hypotheses for the population proportion needs to be tested:

$$H_0 : p_1 = p_2$$
$$H_1 : p_1 \neq p_2$$

This corresponds to a two-tailed test, and a z-test for two population proportions will be used. Based on the information provided, the significance level is $\alpha = 0.05$, and the critical value for a two-tailed test is $z_c = 1.96$.

The rejection region for this two-tailed test is $R = \{z : |z| > 1.96\}$
The rejection region for this two-tailed test is $R = \{z : |z| > 1.96\}$ The z-statistic is computed as follows:

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\bar{p}(1-\bar{p})\left(\frac{1}{N_1} + \frac{1}{N_2}\right)}}$$
$$= \frac{0.75 - 0.5}{\sqrt{0.56(1-0.56)\left(\frac{1}{600} + \frac{1}{900}\right)}} \approx 9.69$$

Since it is observed that $|z| = 6.38 > 1.96 = z_c$, it is then concluded that the null hypothesis is rejected.

Using the P-value approach: The p-value is $p = 2P(Z > 9.682) = 0$, and since $p = 0 < 0.05 = \alpha$, it is concluded that the null hypothesis is rejected.

Therefore, there is enough evidence to claim that the population proportion p_1 is different than p_2, at the $\alpha = 0.05$ significance level.

Problem 78. *A sample of 900 members is found to have a mean of* 3.4 cm. *Can it be reasonably regarded as a truly random sample from a large population with mean* 3.25 cm *and* $S.D.1.61$ cm.

Solution : Here $\bar{x} = 3.4$ cm, $n = 900$, $\mu = 3.25$ and $\sigma = 1.61$ cm $\therefore$

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{3.4 - 3.25}{1.61/30} = 2.8$$

As $z > 1.96$, the deviation of the sample mean from the mean of the population is significant at 5% level of significance. Hence it cannot be regarded as a random sample.

Problem 79. *In an examination given to students at a large number of different schools the mean grade was 74.5 and S.D grade was 8. At one particular school where 200 students took the examination the mean grade was 75.9. Discuss the significance of this result at both 5% and 1% level of significance. (VTU Model 2023)*

Solution :

Problem 80. *One type of air craft is found to develop engine trouble in 5 flights out of a total of 100 and another type in 7 flights out of a total of 200 flights. Is there a significance difference in the two types of air craft's so far as engine defects are concerned? Test at 5% significance level. (VTU Model 2023)*

Solution :

$$p_1 = \frac{5}{100} = 0.05, \quad p_2 = \frac{7}{200} = 0.035$$

$$H_0 : p_1 = p_2$$

$$H_1 : p_1 \neq p_2$$

$$P = \frac{5+7}{100+200} = 0.04, \quad Q = 1 - P = 0.96$$

$$z = \frac{p_1 - p_2}{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

$$= -3.953 (\text{after simplification})$$

$|z| > 1.96$

Hence H_0 is rejected.

Problem 81. *The means of simple samples of sizes 1000 and 2000 are 67.5 and* 68.0 cm *respectively. Can the samples be regarded as drawn from the same population of S.D.* 2.5 cm.

Solution : We have

$$\bar{x}_1 = 67.5, \bar{x}_2 = 68.0$$

$$n_1 = 1000, n_2 = 2000.$$

On the hypothesis, that the samples are drawn from the same population of S.D. $\sigma = 2.5$, we get

$$z = \frac{\bar{x}_1 \sim \bar{x}_2}{\sigma\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{67.5 \sim 68.0}{2.5\sqrt{\left(\frac{1}{1000} + \frac{1}{2000}\right)}}$$

$$= \frac{0.5}{2.5 \times 0.0387} = \frac{0.5}{0.09675} = 5.1$$

Hence the difference between the sample means i.e., 5.1 is very much greater than 1.96 and is therefore significant. Thus, the samples cannot be regarded as drawn from the same population.

Problem 82. *A sample of height of 6400 soldiers has a mean of 67.85 inches and a standard deviation of 2.56 inches while a simple sample of heights of 1600 sailors has a mean of 68.55 inches and a standard deviation of 2.52 inches. Do the data indicate that the sailors are on the average taller than soldiers?*

Solution : Here

$$\bar{x}_1 = 67.85, \sigma_1 = 2.56, n_1 = 6400$$

$$\bar{x}_2 = 68.55, \sigma_2 = 2.52, n_2 = 1600.$$

$\therefore$ S.E. of the difference of the mean heights is

$$e = \sqrt{\left[\frac{{\sigma_1}^2}{n_1} + \frac{\sigma_2^2}{n_2}\right]} = \sqrt{\left[\frac{(2.56)^2}{6400} + \frac{(2.52)^2}{1600}\right]}$$

$$= \sqrt{[.001024 + .003969]} = 0.005 \text{ nearly.}$$

Also difference between the means $= \bar{x}_2 - \bar{x}_1 = 0.7$, which $> 10e$. This is highly significant. Hence the data indicates that the sailors are on the average taller than the soldiers.

3.13 Confidence limits

95% confidence limits for the mean of the population corresponding to a given sample is

$$\bar{x} \pm 1.96(\sigma/\sqrt{n})$$

and 99% confidence limits for the mean is

$$x \pm 2.58(\sigma/\sqrt{n})$$

. 95% confidence limits for the proportion of the population corresponding to a given sample is

$$p \pm 1.96\sqrt{\frac{pq}{n}}$$

99% confidence limits for the proportion of the population corresponding to a given sample is

$$p \pm 2.58\sqrt{\frac{pq}{n}}$$

Problem 83. *A sample of 900 days was taken in a coastal town and it was found that on 100 days the weather was very hot. Obtain the probable limits of the percentage of very hot weather.*

Solution : Probability of very hot weather $= p = \frac{100}{900} = \frac{1}{9} \therefore q = \frac{8}{9}$

Probable limits $= p \pm 2.58\sqrt{pq/n}$

$$= 0.111 \pm (2.58)\sqrt{\frac{1}{9} \times \frac{8}{9} \times \frac{1}{900}} = 0.111 \pm 0.027$$

$$= 0.084 \text{ and } 0.138$$

Problem 84. *In a sample of* 500**men** *it was found that* 60% *of them had over weight. What can we infer about the proportion of people having over weight in the population?*

Solution : Probability of persons having over weight $= p = \frac{60}{100} = 0.6 \quad \&. q = 1 - p = 0.4$
Probable limits $= p \pm 2.58\sqrt{pq/n}$
Probable limits $= 0.6 \pm 2.58\sqrt{\frac{(0.6)(0.4)}{500}}$
Probable limits $= 0.6 \pm 0.0565 = 0.5435$ and 0.6565
Thus the probable limits of people having over weight is 54.35% to 65.65%

Problem 85. *To know the mean weights of all 10 year old boys in Delhi a sample of 225 was taken. The mean weight of the sample was found to be 67 pounds with S.D of 12 pounds. What can we infer about the mean weight of the population?*

Solution: Sample mean $(\bar{x}) = 67$, Sample size $n = 225$, S.D $(\sigma) = 12$
95% confidence limits for the mean of the population corresponding to a given sample is $\bar{x} \pm 1.96(\sigma/\sqrt{n})$
and 99% confidence limits for the mean is $x \pm 2.58(\sigma/\sqrt{n})$.
We have $\cdot \sigma/\sqrt{n} = 12/15 = 0.8$
95% confidence limits : $67 \pm 1.96(0.8) = 65.432$ and 68.568
99% confidence limits : $67 \pm 2.58(0.8) = 64.936$ and 69.064
We can say with 95% confidence that the mean weight of the population lies between 65.4 pounds and 68.6 pounds. Also with 99% confidence we can say that the mean weight lies between 64.9 pounds to 69.1 pounds.

Problem 86. *A sample of* 100**days** *is taken from meteorological records of a certain district and 10 of them are found to he foggy. What are the probable limits of the percentage of foggy days in the district.*

Solution: p = proportion of foggy days in a sample of 100 days is given by $10/100 = 0.1$
Hence $q = 1 - p = 0.9 \therefore$ probable limits of foggy days

$$= p \pm 2.58\sqrt{pq/n}$$
$$= 0.1 \pm 2.58\sqrt{(0.1 \times 0.9)/100}$$
$$= 0.1 \pm 0.0774 = 0.0226 \text{ and } 0.1774$$
$$= 2.26\% \text{ and } 17.74\%$$

Thus the percentage of foggy days lies between 2.26 and 17.74

Problem 87. *A survey was conducted in a slum locality of 2000 families by selecting a sample of size 800. It was revealed that 180 families were illiterates. Find the probable limits of the illiterate families in the population of 2000.*

Solution : Probability of illiterate families $:= p = \frac{180}{800} = 0.225 \therefore q = 0.775$
Probable limits of illiterate families $= p \pm (2.58)\sqrt{pq/n}$

$$\text{i.e.,} \quad = 0.225 \pm (2.58) \uparrow \sqrt{\frac{(0.225)(0.775)}{800}} = 0.225 \pm 0.038$$

$$= 0.187 \text{ and } 0.263 \quad .$$

$\therefore$ the probable limits of illiterate families in the population of 2000 is (0.187)2000 and (0.263) 2000 .
Thus 374 to 526 are probably illiterate families.

Problem 88. *A random sample of 500 apples was taken from a large consignment and 65 were found to be bad. Estimate the proportion of bad apples in the consignment as well as the standard error of the estimate. Also find the percentage of bad apples in the consignment.*

Solution : $p =$ proportion of bad apples in the sample is given by $65/500 = 0.13$.
Hence $q = 1 - p = 0.87$
S.E proportion of bad apples $= \sqrt{pq/n} = \sqrt{(0.13 \times 0.87)/500} = 0.015$
Probable limits of bad apples in the consignment

$$= p \pm 2.58\sqrt{pq/n}$$
$$= 0.13 \pm 2.58(0.015) = 0.13 \pm 0.0387$$
$$= 0.0913 \text{ and } 0.1687$$
$$= 9.13\% \text{ and } 16.87\%$$

Thus the required percentage of bad apples in the consignment lies between 9.13 and 16.87

Problem 89. *In a locality of 18000 families a sample of 840 families was selected at random. Of these 840 families, 206 families were found to have monthly income of Rs. 2500 or less. It was desired to estimate how many of the 18,000 families have monthly income of Rs. 2500 or less. Within what limits would you place your estimate.*

Solution : Proportion of families having monthly income of Rs. 2500 or less is given by

$$p = 206/840 = 0.245. \text{ Hence } q = 1 - p = 0.755$$

S.E proportion $= \sqrt{pq/n} = \sqrt{(0.245 \times 0.755)/840} = 0.015$
Probable limits of families having monthly income of Rs. 2500 or less are $p \pm 2.58\sqrt{pq/n}$. That is,

$$= 0.245 \pm (2.58)(0.015)$$
$$= 0.245 \pm 0.0387$$
$$= 0.2063 \text{ and } 0.2837 \text{ or } 20.63\% \text{ and } 28.37\%$$

Hence the probable limits in respect of 18,000 families is given by

$$0.2063 \times 18,000 \text{ and } 0.2837 \times 18,000$$

That is 3713.4 and 5106.6 or 3713 and 5107
Thus we say that 3713 to 5107 families are likely to have monthly income of Rs. 2500 or less.

Problem 90. *The mean and S.D of the maximum loads supported by 60 cables are 11.09 tonnes and 0.73 tonnes respectively. Find (a)* 95% *(b)* 99% *confidence limits for mean of the maximum loads of all cables produced by the company.*

Solution: By data $\bar{x} = 11.09, \sigma = 0.73$ (a) 95% confidence limits for the mean of maximum loads are given by

$$\bar{x} \pm 1.96(\sigma/\sqrt{n})$$
$$= 11.09 \pm 1.96(0.73/\sqrt{60})$$
$$= 11.09 \pm 0.18$$

Thus 10.91 tonnes to 11.27 tonnes are the 95% confidence limits for the mean of maximum loads. (b) 99% confidence limits for the mean of maximum loads are given by

$$\bar{x} \pm 2.58(\sigma/\sqrt{n})$$
$$= 11.09 \pm 2.58(0.73/\sqrt{60})$$
$$= 11.09 \pm 0.24 = 10.85 \text{ and } 11.33$$

Thus 10.85 tonnes to 11.33 tonnes are the 99% confidence limits for the mean of maximum loads.

Problem 91. *A coin was tossed 400 times and the head turned up 216 times. Test the hypothesis that coin is unbiased at 5% level of significance. (VTU Model 2023 July 2013, 2007)*

Solution: First we will set up null and alternate hypotheses.
Let H_0 : The coin is unbiased.

and H_1 : The coin is biased.
Let x= no. of Heads
As the coin is assumed to be fair, p=P(success)=$\frac{1}{2}$ in each toss.
Hence $q = 1 - p = \frac{1}{2}$
Here coin is tossed $n = 400$ times,
But the observed no. of heads, $x = 216$.
Further, standard error is $e = \sqrt{npq} = \sqrt{(400) \times \frac{1}{2} \times \frac{1}{2}} = 10$

$$z = \frac{x - np}{e} = \frac{216 - 200}{10} = 1.6$$

If we choose the significance level $\alpha = 5\%$, then the tabulated value is 1.96. since, the calculated value is less than the tabulated value; we accept the null hypothesis that coin is un - biased.

Problem 92. *A die was thrown 9000 times and a throw of 5 or 6 was obtained 3240 times. On the assumption of random throwing, do the data indicate an unbiased die? [VTU Model 2023 , Model 2020, Jan 2020, Jan 2018, 2010]*

Solution: We set up the null hypothesis as H_0 : Die is un - biased. and
H_1 : Die is biased.
Let us take level of significance as $\alpha = 5\%$.
Based on the assumption that distribution is normally distributed, the tabulated value is 1.96
The chance of getting each of the 6 numbers is same and it equals to $1/6$ therefore chance of getting either 5 or 6 is $p = \frac{2}{6} = \frac{1}{3}$, $q = 1 - p = \frac{2}{3}$
In a throw of $n = 9000$ times, expected number of number of times to get either 5 or 6 is $np = \frac{1}{3} \times 9000 = 3000$.
Observed no. of successes is $x = 3240$.
$S.E. = \sqrt{npq} = 44.72$.
Now consider the test criterion is
$z = \frac{x-np}{\sqrt{npq}} = \frac{3240}{44.72} = 5.367$
Here $|z| > z_{0.05} = 1.96$
Therefore, we reject null hypothesis and accept the alternate hypothesis that die is highly biased.

Problem 93. *In a locality containing 18000 families, a sample of 840 families was selected at random. Of these 840 families, 206 families were found to have a monthly*

income of $|250$ *or less. It is desired to estimate how many out of 18,000 families have a monthly income of* $|250$ *or less. Within what limits would you place your estimate?*

Solution: Here

$$p = \frac{206}{840} = \frac{103}{420} \text{ and } q = \frac{317}{420}$$

$\therefore$ standard error of the population of families having a monthly income of $|250$ or less

$$= \sqrt{\left(\frac{pq}{n}\right)} = \sqrt{\left(\frac{103}{420} \times \frac{317}{420} \times \frac{1}{840}\right)} = .015 = 1.5\%$$

Hence taking $\frac{103}{420}$ (or 24.5%) to be the estimate of families having a monthly income of | 250 or less in the locality, the limits are $(24.5 \pm 3 \times 1.5)\%$ i.e., 20% and 29% approximately.

Question Bank

1. A die is tossed 960 times and it falls with 5 upwards 184 times. Is the die biased?
2. A coin is tossed 1000 times and head turns up 540 times. Test the hypothesis that the coin is an unbiased one.
3. It is claimed that a random sample of 49 tyres has a mean life of $15,200$kms. Is the sample drawn from a population whose mean is $15,150$kms and whose standard deviation is 1,200 kms? Test the significance at 0.05 level. (VTU Model 2020)
4. A sample of 900 members is found to have a mean of 3.4 cm. Can it be reasonably regarded as a truly random sample from a large population with mean 3.25 cm and S.D. 1.61 cm.
5. In an examination given to students at a large number of different schools the mean grade was 74.5 and S.D grade was 8. At one particular school where 200 students took the examination the mean grade was 75.9. Discuss the significance of this result at both 5% and 1% level of significance. (VTU Model 2023) Ans: z=2.4748
6. One type of air craft is found to develop engine trouble in 5 flights out of a total of 100 and another type in 7 flights out of a total of 200 flights. Is there a

significance difference in the two types of air craft's so far as engine defects are concerned? Test at 5% significance level. (VTU Model 2023)

7. In a sample of 600 men from a certain city, 450 are found smokers. In another sample of 900 men from another city, 450 are smokers. Do the data indicate that the cities are significantly different with respect to the habit of smoking among men. Test at 5% significance level.
   z=6.38

8. In a large city A, 20% of a random sample of 900 school boys had a slight physical defect. In another large city B, 18.5% of a random sample of 1600 school boys had the same defect. Is the difference between the proportions significant at 5% significance level? (VTU Model 2023)

9. A sample of 100 tyres is taken from a lot. The mean life of a tyre is found to be 39350 kms with a SD of 3260. Can it be considered as the true random sample from population with mean life of 40000 kms? (use 5% significance level).

10. In two large populations there are $\mathbf{30\%}$ and $\mathbf{25\%}$ respectively of fair haired people. Is this difference likely to be hidden in samples of 1200 and 900 respectively from the two populations?

11. A stenographer claims that she can type at the rate of 120 words per minute. Can we reject her claim on the basis of 100 trails in which she demonstrates a mean of 116 words with a standard deviation of 15 words? Use 5% level of significance.

12. 12 dice are thrown 3086 times and a throw of $\mathbf{2, 3, 4}$ is reckoned as a success. Suppose that 19142 throws of $\mathbf{2, 3, 4}$ have been made out. Do you think that this observed value deviates from the expected value? If so, can the deviation from the expected value be due to fluctuations of simple sampling?

13. A sample of 100 students is taken from a large population. The mean height of the students in this sample is 160cm. Can it be reasonably regarded that this sample is from a population of mean 165 cm and S.D 10 cm?

14. A die is tossed 960 times and 5 appear 184 times, is the die biased?

15. In 324 throws of a six faced 'die' an odd number turned up 181 times. Is it reasonable to think that the die is an unbiased one at 1% level of significance.? Ans: z=2.11, accept

16. In a sample of 600 men from a certain city, 450 are found smokers. In another sample of 900 men from another city, 450 are smokers. Do the data indicate that the cities are significantly different with respect to the habit of smoking among men. Test at 5% significance level.

17. The means of simple samples of sizes 1000 and 2000 are 67.5 and **68.0 cm** respectively. Can the samples be regarded as drawn from the same population of S.D. **2.5 cm**.
Ans : z=5.1

18. A sample of height of 6400 soldiers has a mean of 67.85 inches and a standard deviation of 2.56 inches while a simple sample of heights of 1600 sailors has a mean of 68.55 inches and a standard deviation of 2.52 inches. Do the data indicate that the sailors are on the average taller than soldiers?
Ans : reject

19. In a group of 50 first cousines there were found to be 27 males and 23 females. Ascertain if the observed proportions are inconsistent with the hypothesis that the sexes should be in equal proportion.

20. A random sample of 500 apples was taken from a large consignment and 65 were found to be bad. Estimate the proportion of the bad apples in the consignment as well as the standard error of the estimate. Deduce that the percentage of bad apples in the consignment almost certainly lies between 8.5 and 17.5.

21. The mean life time of a sample of 100 fluorescent tube lights manufactured by a company is found to be 1570 hrs with a standard deviation of 120 hrs. Test the hypothesis that the mean life-time of the lights produced by the company is 1600 hrs at 0.01 level of significance.

22. A machine produces 16 imperfect articles in a sample of 500. After machine is overhauled, it produces 3 imperfect articles in a batch of 100 . Has the machine been improved? Ans: z=0.1047, accept

23. n a sample of 500 people from a state 280 take tea and rest take coffee. Can we assume that tea and coffee are equally popular in the state at 5% level of significance? Ans: 2.68, reject

24. A sample of 400 items is taken from a normal population whose mean is 4 and variance 4 . If the sample mean is 4.45 , can the samples be regarded as a simple sample?

25. To know the mean weights of all 10-year old boys in Delhi, a sample of 225 is taken. The mean weight of the sample is found to be 67 pounds with a S.D. of 12 pounds. Can you draw any inference from it about the mean weight of the population?

26. A normal population has a mean 0.1 and a S.D. of 2.1. Find the probability that the mean of simple sample of 900 members will be negative.

27. If the mean breaking strength of copper wire is $\mathbf{575lbs}$. with a standard deviation of $\mathbf{8.3lbs}$, how large a sample must be used in order that there be one chance in 100 that the mean breaking strength of the sample is less than 572 lbs.? [Hint. $|z| = \left|\frac{\bar{x}-\mu}{\sigma}\sqrt{n}\right| = \frac{3}{8.3}\sqrt{n}$ Also from table IV, $z = \mathbf{2.33}$. Hence $n = \mathbf{42}$ nearly.]

28. A research worker wishes to estimate mean of a population by using sufficiently large sample. The probability is $\mathbf{95\%}$ that sample mean will not differ from the true mean by more than $\mathbf{25\%}$ of the S.D. How large a sample should be taken?

29. A machine produces 16 imperfect articles in a sample of 500 . After machine is overhauled, it produces 3 imperfect articles in a batch of 100 . Has the machine been improved?

30. A sample of 400 items is taken from a normal population whose mean is 4 and variance 4 . If the sample mean is 4.45 , can the samples be regarded as a simple sample?

31. A sample of 100 electric bulbs produced by manufacturer $\boldsymbol{A}$ showed a mean life time of 1190 hours and a standard deviation of 90 hours. A sample of 75 bulbs produced by manufacturer $\boldsymbol{B}$ showed a mean life time of 1230 hours, with a standard deviation of 120 hours. Is there a difference between the mean life time of two brands at a significance level of $(i)\mathbf{0.05}$ (ii) 0.01 .

32. A random sample of 1000 men from North India shows that their mean wage is $|\mathbf{5}$ per day with a S.D. of $|\mathbf{1.50}$. A sample of 1500 men from South India gives a mean wage of $|\mathbf{4.50}$ per day with a standard deviation of $|\mathbf{2}$. Does the mean rate of wages varies as between the two regions?

33. If a sample of 300 units of a manufactured product 65 units were found to be defective and in another sample of 200 units, there were 35 defectives. Is there significant difference in the proportion of defectives in the samples at 5% Level. Ans : z=1.233, accept

34. A survey was conducted in a slum locality of 2000 families by selecting a sample of size 800. It was revealed that 180 families were illiterates. Find the probable limits of the illiterate families in the population of 2000.

35. To know the mean weights of all 10 year old boys in Delhi a sample of 225 was taken. The mean weight of the sample was found to be 67 pounds with S.D of 12 pounds. What can we infer about the mean weight of the population?

36. A sample of 900 days was taken in a coastal town and it was found that on 100 days the weather was very hot. Obtain the probable limits of the percentage of very hot weather.

37. In a sample of $\mathbf{500men}$ it was found that $\mathbf{60\%}$ of them had over weight. What can we infer about the proportion of people having over weight in the population?

38. In a locality of 18000 families a sample of 840 families was selected at random. Of these 840 families, 206 families were found to have monthly income of Rs. 2500 or less. It was desired to estimate how many of the 18,000 families have monthly income of Rs. 2500 or less. Within what limits would you place your estimate.

39. The mean and S.D of the maximum loads supported by 60 cables are 11.09 tonnes and 0.73 tonnes respectively. Find (a) $\mathbf{95\%}$ (b) $\mathbf{99\%}$ confidence limits for mean of the maximum loads of all cables produced by the company.

40. 400 children are chosen in an industrial town and 150 are found to be under weight. Assuming the conditions of simple sampling, estimate the percentage of children who are under weight in the industrial town and assign limits within which the percentage probably lies?

Module 4

Statistical Inference II

Syllabus :

Sampling variables, central limit theorem and confidences limit for unknown mean, Test of Significance for means of two small samples, students 't'distribution, Chi-square distribution as a test of goodness of fit. F-Distribution.

4.1 Central Limit Theorem:

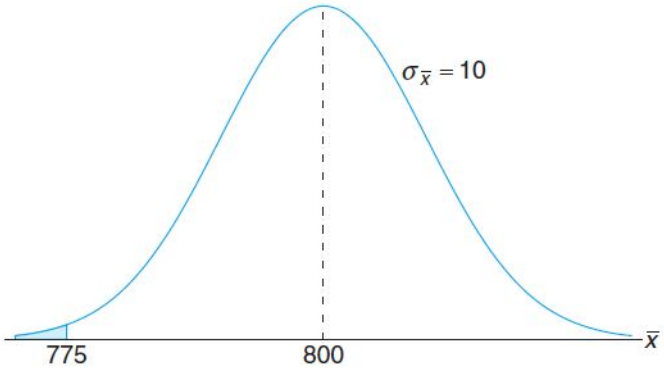
If $\bar{X}$ is the mean of a random sample of size n taken from a population with mean μ and finite variance σ^2, then the limiting form of the distribution of

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}},$$

as $n \to \infty$, is the standard normal distribution (with mean 0 and S.D. 1).
In a population whose distribution may be known or unknown, if the size (n) of samples is sufficiently large, the distribution of the sample means will be approximately normal.

Problem 94. *An electrical firm manufactures light bulbs that have a length of life that is approximately normally distributed, with mean equal to 800 hours and a standard deviation of 40 hours. Find the probability that a random sample of 16 bulbs will have an average life of less than 775 hours.*

Solution: The sampling distribution of $\bar{X}$ will be approximately normal, with $\mu_{\bar{X}} = 800$ and $\sigma_{\bar{X}} = 40/\sqrt{16} = 10$. The desired probability is given by the area of the shaded region inthe following figure.



$$
\begin{aligned}
P(\bar{X} < 775) &= P\left(\frac{\bar{X} - \mu}{\frac{sigma}{\sqrt{n}}} < \frac{775 - \mu}{\frac{sigma}{\sqrt{n}}}\right) \\
&= P(Z < -2.5) \\
&= Area(-\infty to - 2.5) \\
&= Area(2.5to\infty) \\
&= A(\infty) - A(2.5) \\
&= 0.0062
\end{aligned}
$$

Problem 95. *An unknown distribution has a mean of 90 and a standard deviation of 15 . A sample size of 80 is drawn randomly from the population. Find the probability that the sum of the 80 values (or the total of the 80 values) is more than 7400 .*

Solution: When the sum of the 80 values is 7400 , the sample mean is given as:

$$
\begin{aligned}
\bar{x} &= \frac{\sum x}{n} \\
&= \frac{7400}{80} \\
&= 92.5
\end{aligned}
$$

The probability that the sum of the 80 values is more than 7400 or the probability that the sample mean of 80 values is more than 92.5 :

$$
\begin{aligned}
P(\bar{x} > 92.5) &= P\left(z > \frac{92.5 - 90}{1.68}\right) \\
&= P(z > 1.49) \\
&= P(z < -1.49) \\
&= 0.0681
\end{aligned}
$$

Problem 96. *An unknown distribution has a mean of 90 and a standard deviation of 15 . Samples of size* $n = 25$ *are drawn randomly from the population. Find the probability that the sample mean is between 85 and 92 .*

Solution: Let $X =$ one value from the original unknown population. The probability question asks you to find a probability for the sample mean.Let $\bar{X} =$ the mean of a sample of size 25. Since $\mu = 90$, $\sigma = 15$, and $n = 25$,

$$
\begin{aligned}
& P(85 < \bar{x} < 92) \\
&= P\left(\frac{85-\mu}{\frac{sigma}{\sqrt{n}}} < \frac{\bar{X}-\mu}{\frac{sigma}{\sqrt{n}}} < \frac{92-\mu}{\frac{sigma}{\sqrt{n}}}\right) \\
&= P\left(\frac{85-90}{15/\sqrt{25}} < z < \frac{92-90}{15/\sqrt{25}}\right) \\
&= P(-1.67 < z < 0.67) \\
&= \text{Area under the standard normal curve from } -1.67 \text{ to } 0.67 \\
&= \text{Area under the standard normal curve from } -1.67 \text{ to } 0 \\
&\quad + \text{Area under the standard normal curve from } 0 \text{ to } 0.67 \\
&= \text{Area under the standard normal curve from } 0 \text{ to } 1.67 \\
&\quad + \text{Area under the standard normal curve from } 0 \text{ to } 0.67 \\
&= A(1.67) + A(0.67) \\
&= 0.6997
\end{aligned}
$$

Problem 97. *State Central limit theorem. Use the theorem to evaluate* $P[50 < \bar{X} < 56]$ *where* $\bar{X}$ *represents the mean of a random sample of size 100 from an infinite population with mean* $\mu = 53$ *and variance* $\sigma^2 = 400$

Solution: Central limit theorem : If $\bar{X}$ is the mean of a random sample of size n taken from a population with mean μ and finite variance σ^2, then the limiting form of the distribution of

$$Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}},$$

as $n \to \infty$, is the standard normal distribution (with mean 0 and S.D. 1).
Given sample size, $n = 100$,
population mean $\mu = 53$,
population variance, $\sigma^2 = 400$

Hence $\sigma = 20$

$$P[50 < \bar{X} < 56] = P\left[\frac{50-\mu}{\sigma/\sqrt{n}} < \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < \frac{56-\mu}{\sigma/\sqrt{n}}\right]$$
$$= P(-1.5 < z < 1.5)$$
$$= \text{Area under the standard normal curve from } -1.5 \text{ to } 1.5$$
$$= 2A(1.5) \quad \text{(by symmetry)}$$
$$= 2 \times 0.4332$$
$$= 0.8664$$

Problem 98. *Certain tubes manufactured by a company have mean life time of 800 hours and S.D of 60 hours. Find the probability that a random sample of 16 tubes taken from the group will have a mean life time (a) between 790 hours and 810 hours. (b) less than 785 hours. (c) more than 820 hours. (d) between 770 hours and 830 hours.*

Solution: By data $\mu = 800, \sigma = 60, n = 16 \therefore \quad \sigma_{\bar{x}} = \sigma/\sqrt{n} = 60/4 = 15$

(a)

$$P(790 < \bar{x} < 810) = P\left[\frac{790-\mu}{\sigma/\sqrt{n}} < \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < \frac{810-\mu}{\sigma/\sqrt{n}}\right]$$
$$= P(-0.67 < z < 0.67)$$
$$= 2P(0 < z < 0.67)$$
$$= 2(0.2486)$$
$$= 0.4972$$

Thus $P(790 < \bar{x} < 810) = 0.4972$

(b)

$$P(\bar{x} < 785) = P\left[\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < \frac{785-\mu}{\sigma/\sqrt{n}}\right]$$
$$= p(z < -1)$$
$$= P(z > 1)$$
$$= A(\infty) - A(1)$$
$$= 0.5 - 0.3413$$
$$= 0.1587$$

Thus $P(\bar{x} < 785) = 0.1587$

(c)

$$\begin{aligned} P(\bar{x} > 820) &= P\left[\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} > \frac{820 - \mu}{\sigma/\sqrt{n}}\right] \\ &= P(z > 1.33) \\ &= P(z > 0) - P(0 < z < 1.33) \\ &= A(\infty) - A(1.33) \\ &= 0.5 - 0.4082 \\ &= 0.0918 \end{aligned}$$

Thus $P(\bar{X} > 820) = 0.0918$

(d)

$$\begin{aligned} P(770 < \bar{x} < 830) &= P\left[\frac{770 - \mu}{\sigma/\sqrt{n}} < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < \frac{830 - \mu}{\sigma/\sqrt{n}}\right] \\ &= P(-2 < z < 2) \\ &= 2P(0 < z < 2) \\ &= 2(0.4772) \\ &= 0.9544 \end{aligned}$$

Thus $P(770 < \bar{x} < 830) = 0.9544$

Problem 99. *A random sample of size* 64 *is taken from an infinite population having mean* 112 *and variance* 144*. Using central limit theorem, find the probability of getting the sample mean* $\bar{X}$ *greater than* 114.5

Solution: Given $n = 64$

$\mu = 112$

$\sigma^2 = 144$

$$\begin{aligned} P(\bar{X} > 114.5) &= P\left[\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} > \frac{114.5 - \mu}{\sigma/\sqrt{n}}\right] \\ &= P\left[z > \frac{114.5 - 112}{1.5}\right] \\ &= P(z > 1.66) \\ &= A(\infty) - A(1.66) \\ &= 0.5 - 0.4515 \\ &= 0.0489 \end{aligned}$$

Problem 100. *In a recent study reported on the Flurry Blog, the mean age of tablet users is 34 years. Suppose the standard deviation is 15 years. Take a sample of size* $n = 100$. *Using central limit theorem, find the probability that the sample mean age is more than 30 years.*

Solution: Given, Sample size $\mathbf{n = 100}$
Mean of the population $\boldsymbol{\mu = 34}$
S.D. of the population $\Rightarrow \boldsymbol{\sigma = 15}$

$$
\begin{aligned}
P(\bar{X} > 30) &= P\left[\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} > \frac{30 - \mu}{\sigma/\sqrt{n}}\right] \\
&= P\left[z > \frac{30 - 34}{15/10}\right] \\
&= P\left[z > -2.66\right] \\
&= \text{Area from -2.66 to } \infty \ (understandardnormalcurve) \\
&= \text{Area from -2.66 to 0} + \text{Area from 0 to } \infty \\
&= \mathbf{A(2.66) + 0.5} \\
&= \mathbf{0.9961}
\end{aligned}
$$

4.2 Sampling of Variables-Small samples

In case of large samples, sampling distribution approaches a normal distribution and values of sample statistic are considered best estimates of the parameters in a population. It will no longer be possible to assume that statistics computed from small samples are normally distributed. As such, a new technique has been devised for small samples which involves the concept of 'degrees of freedom' which we explain below.

Number of degrees of freedom is the number of values in a set which may be assigned arbitrarily. For instance, if $x_1 + x_2 + x_3 = 15$ and we assign any values of two of the variables (say : x_1, x_2), then the values of x_3 will be known. The two variables are therefore, free and independent choices for finding the third. Hence these are the degrees of freedom.
If there are n observations, the degrees of freedom (d.f.) are $(n - 1)$.
In other words, while finding the mean of a small sample, one degree of freedom is used up and $(n - 1)$ d.f. are left to estimate the population variance.

4.3 Testing of hypothesis for small samples : Student's t- test

A sample is called a **small sample** if it contains $\leq$ **30** observations. For small samples we use student's distribution which is bell shaped and symmetrical.

If $x_1, x_2, x_3, \ldots..x_n$ be any random sample of size $n \leq 30$ drawn from a normal or approximately normal population with mean μ and variance σ^2 then the student's t statistic is defined as

$$t = \frac{\bar{x} - \mu}{s}\sqrt{n}$$

where μ is the mean of population,

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{\infty} x_i$$

is the mean of the sample and

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x - \bar{x})^2$$

is the variance of the sample.

We compute the test-statistic under H_0 and compare it with the tabulated value of t for $(n-1)$ d.f. at the given level of significance.

If calculated $|t| < t_\alpha$, then H_0 is accepted and we say that difference between sample mean and population mean is not significant at level α.

Problem 101. *A random sample of 10 boys had the following IQ :* $70, 120, 110, 101, 88, 83, 95, 98.107, 100$. *Do these data support the assumption of a population mean* IQ *of* 100*? (at* 5% *level of significance? [VTU Model 2020,2006]*

Solution: Let H_0 : population mean IQ is $\mu = 100$

Given the I.Q. of 10 boys

$$x : 70, 120, 110, 101, 88, 83, 95, 98, 107, 100$$

$$\therefore \bar{x} = \frac{1}{10}\sum x = \frac{972}{10} = 97.2$$

x	70	120	110	101	88	83
$(x-\bar{x})^2$	739.84	519.84	163.84	14.44	84.64	201.64

95	98	107	100	$\sum x = 972$
4.84	0.64	96.04	7.84	$\sum(x-\bar{x})^2 = 1833.6$

Variance
$s^2 = \frac{1}{n-1}\sum(x-\bar{x})^2$
$\Rightarrow s^2 = \frac{1}{9} \times 1833.6$
$\Rightarrow s^2 = 203.73333$
$\Rightarrow s = 14.2735$
Given the mean of population $\mu = 100$ We have
$t = \frac{\bar{x}-\mu}{s}\sqrt{n}$
$\Rightarrow t = \frac{97.2-100}{14.2735}\sqrt{n}$
$\Rightarrow t = \frac{-2.8}{14.2735}\sqrt{10} = -0.6203 < t_{0.05} = 2.262$
Hence we accept H_0. i.e. the data supports the assumption of a population mean IQ= 100

Problem 102. *A certain stimulus administered to each of 12 patients resulted in the following increase of blood pressure :* $5, 2, 8, -1, 3, 0, -2, 1, 5, 0, 4, 6$. *Can it be concluded that the stimulus will in general be accompanied by an increase in blood pressure?* *[VTU Model* **2020**, *June* **2019**, *Jan* **2018**, *Dec* **2012**, *Dec* **2010, 2007***]*

Solution: H_0 : Assume that the stimulus will not be accompanied by increase in blood pressure. Under this assumption, the mean of increase in blood pressure for the population is zero, we have $\mu = 0$

$$\bar{x} = \frac{5+2+8-1+3+0-2+1+5+0+4+6}{12}$$

$$= \frac{31}{12} = 2.583 = 2.6 \text{ approx.}$$

x	$x-2.6$	$(x-\bar{x})^2=(x-2.6)^2$
5	2.4	5.76
2	−0.6	0.36
8	5.4	29.16
−1	−3.6	12.96
3	0.4	0.16
0	−2.6	6.76
−2	−4.6	21.16
1	−1.6	2.56
5	2.4	5.76
0	−2.6	6.76
4	1.4	1.96
6	3.4	11.56
$\Sigma x=31$		$\Sigma(x-\bar{x})^2=104.92$

$$s^2=\frac{\sum(x-\bar{x})^2}{n-1}=\frac{104.92}{12-1}=9.54$$

$$s=3.08$$

The t statistic is,

$$t=\frac{\bar{x}-\mu}{s}\sqrt{n}=\left(\frac{2.6-0}{3.08}\right)\sqrt{12}=\frac{2.6}{3.08}\times 3.464=2.924$$

As the computed value of t, i.e., 2.924 is greater than $t_{0.05}=2.201$ for 11 d.f. Hence we reject H_0 and we conclude that as a result of the stimulus blood pressure will increase.

Problem 103. *Ten individuals are chosen at random from the population and their heights are found to be inches* $63, 63, 64, 65, 66, 69, 69, 70, 70, 71$. *Discuss the suggestion that the mean height in the universe is* 65 *inches, given that for* 9 *degree of freedom the value of student's 't' at* 0.05 *level of significance is* 2.262. *[VTU: Dec 2018, DEC/JAN 16, Jan 2014, June 2012]*

Solution: The null hypothesis is $\mathrm{H}_0: \mu=65$ inches

The alternative hypothesis $\mathrm{H}_\alpha: \mu\neq 65$ inches

The $\mathrm{d.f}, \nu=\mathrm{n}-1=10-1=9$

$x_i=63, 63, 64, 65, 66, 69, 69, 70, 70, 71$ and $n=10$

$\therefore \bar{x}=\frac{\Sigma x_i}{n}=\frac{670}{10}=67$

and $(x_i-\bar{x})^2$ are given by $(63-67)^2=16$, $(63-67)^2=16$, $(64-67)^2=$

$9,\ (65-67)^2=4,\ (66-67)^2=1,\ (69-67)^2=4,\ (69-67)^2=4,$
$(70-67)^2=9,\ (70-67)^2=9,\ (71-67)^2=16,$
$s^2=\frac{1}{n-1}\sum_{i=1}^{n}(x-\bar{x})^2$
$s=\sqrt{\frac{\Sigma(x_i-\bar{x})^2}{n-1}}=\sqrt{\frac{88}{9}}=3.13$

$$t=\frac{|\bar{x}-\mu|}{s}\sqrt{n}=\frac{67-65}{3.13}\sqrt{10}=2.02$$

But $t_{0.05}=2.262$ at 9 d.f.(given)
i.e.. $|t|=2.02<t_{0.05}=2.262$
The difference is not significant at a 0.05 level and Ho is accepted and we conclude that the mean height is 65 inches.

Problem 104. *Nine items have values* $45, 47, 50, 52, 48, 47, 49, 53, 51$. *Does the mean of these differ significantly from assumed mean of* 47.5*?(*$\nu=8, t_{0.05}=2.31$*)* *[VTU:JUNE/JULY-15, July 2013, 2010]*

Solution: $\mathbf{H_0}: \mu=47.5$ i.e. there is no significant difference between the sample and population mean.
$\mathbf{H_1}: \mu\neq 47.5$; Given $: n=9,$

x	45	47	50	52	48	47	49	53	51
$x-\bar{x}$	-4.11	-2.11	0.89	2.89	-1.11	-2.11	-0.11	3.89	1.89
$(x-\bar{x})^2$	16.89	4.45	0.79	8.35	1.23	4.45	0.01	15.13	3.57

$\bar{x}=\frac{\Sigma x}{n}=\frac{442}{9}=49.11;$
$\Sigma(x-\bar{x})^2=54.89$;
$s^2=\frac{\Sigma(x-\bar{x})^2}{(n-1)}=6.86 \quad \therefore \quad s=2.6192$

$$t=\frac{\bar{x}-\mu}{s/\sqrt{n}}=\frac{49.11-47.5}{2.6192/\sqrt{9}}=1.84$$

But $t_{0.05}=2.31$ for $\nu=8$
Conclusion : since $|t|<t_{0.05}$, the hypothesis is accepted.
i.e. there is no significant difference between their mean.

Problem 105. *A machinist is making engine parts with axle diameter of* 0.7 *inch. A random sample of* 10 *parts shows mean diameter of* 0.742 *inch with a standard deviation of* 0.04 *inch. On the basis of this sample, would you say that the work is inferior?* *[VTU 2009]*

Solution : Let H_0: product is not inferior $i.e.$, there is no significant difference between $\bar{x}$ and μ

Here we have $\mu = 0.700, \bar{x} = 0.742, s = 0.040, n = 10$

$\therefore$

$$t = \frac{\bar{x} - \mu}{s}\sqrt{n} = \frac{0.742 - 0.700}{0.040}\sqrt{10} = \frac{0.133}{0.040} = 3.32$$

Degrees of freedom $\nu = 10 - 1 = 9$

For $\nu = 9$, tabulated value is $t_{0.05} = 2.262$.

As the calculated value of $|t| > t_{0.05}$, we reject H_0. and conclude that the work is inferior.

4.4 Confidence Limits

The confidence limits for the mean of the population corresponding to a given **large sample** (i.e. $n \geq 30$) is

$$\bar{x} \pm z_\alpha(\sigma/\sqrt{n})$$

where σ is the population S.D. and n is the sample size.

In particular, 95% confidence limits for the mean of the population corresponding to a given sample is

$$\bar{x} \pm 1.96(\sigma/\sqrt{n})$$

and 99% confidence limits for the mean is

$$x \pm 2.58(\sigma/\sqrt{n})$$

. The confidence limits for the mean of the population corresponding to a given **small sample** is

$$\bar{x} \pm t_\alpha(s/\sqrt{n})$$

where s is the sample S.D. and n is the sample size and t_α is the table value of t for level α and $n - 1$ degrees of freedom.

Problem 106. *A random sample of size 25 from a normal distribution* $N\left(\mu, \sigma^2 = 4\right)$ *yields, sample mean* $\bar{X} = 78.3$. *Obtain a* 99% *confidence interval for* μ.

Solution: Here sample mean, $\bar{X} = 78.3$,

population mean, $\sigma^2 = 4$,

Sample size, $n = 25$ (small sample) For 99% confidence, we have level of significance $\alpha = 1\% = 0.01$,

For this level, the table value of t is $t_{0.01} = 2.797$

$$\text{confidence interval} = (\bar{x} - t_\alpha \frac{s}{\sqrt{n}}, \bar{x} + t_\alpha \frac{s}{\sqrt{n}})$$
$$= \left(78.3 - 2.797 \times \frac{2}{5}, 78.3 + 2.797 \times \frac{2}{5}\right)$$
$$= (78.3 - 1.1188, 78.3 + 1.1188)$$
$$= (77.1812, 79.4188)$$

Hence 99% confidence confidence interval for μ is $(77.1812, 79.4188)$.

Problem 107. *The heights of a random sample of 50 college students showed a mean of 174.5 centimeters and a standard deviation of 6.9 centimeters. Construct a* 98% *confidence interval for the mean height of all college students.*

Solution: Here $n = 50$ (large)
and $\bar{x} = 174.5$, $s = 6.9$
Since n is large, we shall use z- distribtion.
For 98% confidence, we have level of significance $\alpha = 2\% = 0.02$,
For this level, the table value of z is $z_{0.02} = 2.326$(refer standard normal table)

$$\text{confidence interval} = \left(x - z \times \frac{\sigma}{\sqrt{n}}, x + z \times \frac{\sigma}{\sqrt{n}}\right)$$
$$= \left(174.5 - 2.326 \times \frac{6.9}{\sqrt{50}}, 174.5 + 2.326 \times \frac{6.9}{\sqrt{50}}\right)$$
$$= (172.24, 176.76)$$

Problem 108. *Deduce that for a random sample of 16 values with mean 41.5 inches and the sum of the squares of the deviations from the mean 135 inches* ^{2} *and drawn from a normal population,* 95% *confidence limits for the mean of the population are 39.9 and 43.1 inches.*

Solution: Here, $n = 16$ Hence degrees of freedom, $v = n - 1 = 15$, $\bar{x} = 41.5$
sum of the squares of the deviations from the mean, $\sum(x - \bar{x})^2 = 135$
Hence variance, $s^2 = \frac{\sum(x-\bar{x})^2}{n-1} = \frac{135}{15} = 9$
Sample S.D., $s = \sqrt{9} = 3$
For 95% confidence, we have level of significance $\alpha = 5\% = 0.05$,

For this level, the table value of t for 15 d.f. is is $t_\alpha = t_{0.05} = 2.131$

$$\text{confidence limits} = \bar{x} - t_\alpha \frac{s}{\sqrt{n}} \text{ and } \bar{x} + t_\alpha \frac{s}{\sqrt{n}}$$
$$= 41.5 - 2.131 \times \frac{3}{\sqrt{16}} \text{ and } 41.5 + 2.131 \times \frac{3}{\sqrt{16}}$$
$$= 41.5 - 1.598 \text{ and } 41.5 + 1.598$$
$$= 39.9 \text{ and } 43.1$$

Hence the required confidence limits are 39.9 and 43.1 inches.

Problem 109. *Let the observed value of the mean $\bar{X}$ of a random sample of size 20 from a normal distribution with mean μ and variance $\sigma^2 = 80$ be 81.2 . Find a 90% and a 95% confidence intervals for μ.*

Solution: Here sample mean, $\bar{X} = 81.2$,
population mean, $\sigma^2 = 80 \Rightarrow \sigma = \sqrt{80}$,
Sample size, $n = 20$ (small sample) For 90% confidence, we have level of significance $\alpha = 10\% = 0.1$,
For this level, the table value of t is $t_\alpha = t_{0.1} = 1.729$

$$\text{confidence interval} = (\bar{x} - t_\alpha \frac{\sigma}{\sqrt{n}}, \bar{x} + t_\alpha \frac{\sigma}{\sqrt{n}})$$
$$= \left(81.2 - 1.729 \times \frac{\sqrt{80}}{\sqrt{20}}, 81.2 + 1.729 \times \frac{\sqrt{80}}{\sqrt{20}}\right)$$
$$= (81.2 - 3.458, 81.2 + 3.458)$$
$$= (77.742, 84.658)$$

Hence 90% confidence confidence interval for μ is $(77.742, 84.658)$.

Problem 110. *Suppose scores on exams in statistics are normally distributed with an unknown population mean and a population standard deviation of 3 points. A random sample of 36 scores is taken and gives a sample mean of 68 points. Find a 95% confidence interval for the mean exam score.*

Solution: From the question $\bar{x} = 68, \sigma = 3$ and $n = 36$.
For 95% confidence, $z = 1.96$.
The 95% confidence interval is

$$\left(x - z \times \frac{\sigma}{\sqrt{n}}, x + z \times \frac{\sigma}{\sqrt{n}}\right)$$

$$\text{Lower Limit} = \bar{x} - z \times \frac{\sigma}{\sqrt{n}}$$
$$= 68 - 1.9599\ldots \times \frac{3}{\sqrt{36}}$$
$$= 67.02$$
$$\text{Upper Limit} = \bar{x} + z \times \frac{\sigma}{\sqrt{n}}$$
$$= 68 + 1.9599\ldots \times \frac{3}{\sqrt{36}}$$
$$= 68.98$$

Hence the 95% confidence interval is (67.02, 68.98)

4.5 Test of significance of difference between sample means(small samples)

Consider two independent samples x_i $(i = 1, 2, \cdots n_1)$ and y_j $(j = 1, 2, \cdots n_2)$ drawn from a normal population.
Let $(\bar{x}, s_1)$ and $(\bar{y}, s_2)$ respectively be the mean and variance of the two samples.
Let μ be the population mean and σ be the population variance.
We need to test the hypothesis whether the difference between the sample means is significant.
We compute

$$t = \frac{\bar{x} - \bar{y}}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

where

$$S^2 = \frac{1}{n_1 + n_2 - 2}\left\{\sum_{i=1}^{n_1}(x_i - \bar{x})^2 + \sum_{j=1}^{n_2}(y_j - \bar{y})^2\right\}$$

or

$$s^2 = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}$$

and degrees of freedom $\nu = n_1 + n_2 - 2$ if $|t| > t_{.05}$ the difference between the sample means is said to be significant at 5%. level of significance. (Similarly we can check for 1% level of significance also.)

Problem 111. *Below are given the gain in weights of cows fed on two diets A and B*
Gain in weight:
Diet A: $25, 32, 30, 34, 24, 14, 32, 24, 30, 31, 35, 25$

Diet B: $44, 34, 22, 10, 47, 31, 40, 30, 32, 35, 18, 21, 35, 29, 22$
Test if the two diets differ significantly as regards their effect on increase in weight.

Solution: Setup Null and alternative hypothesis
H_0 : There is no significance of differecne between two sample means
Here

$$\mathbf{n_1 = 12}$$
$$\mathbf{n_2 = 15}$$
$$\sum x_1 = 336$$
$$\sum x_2 = 450$$
$$\sum (x_1 - \bar{x}_1)^2 = 380$$
$$\sum (x_2 - \bar{x}_2)^2 = 1410$$
$$\bar{x}_1 = \frac{1}{n_1}\sum_{i=1}^{n} x_{1i} = \frac{336}{12} = 28$$
$$\bar{x}_2 = \frac{1}{n_2}\sum_{i=1}^{n} x_{2i} = \frac{450}{15} = 30$$

Let the level of significance be 5% with $n_1 + n_2 - 2$ degrees of freedom Tabulated $t_{0.05}$ for $(12 + 15 - 2) = 25$ d. f=2.06

X_1	$x_1 - \bar{x}_1$	$(x_1 - \bar{x}_1)^2$	x_2	$x_2 - \bar{x}_2$	$(x_2 - \bar{x}_2)^2$
25	-3	9	44	14	196
32	4	16	34	4	16
30	2	4	22	-8	64
34	6	36	10	-20	400
24	-4	16	47	17	289
14	-14	196	31	1	1
32	4	16	40	10	100
24	-4	16	30	0	0
30	2	4	32	2	4
31	3	9	35	5	25
35	7	49	18	-12	144
25	-3	9	21	9	81
			35	5	25
			29	-1	1
			22	8	64
$\sum X_1$		$\sum (X_1 - \bar{x}_1)^2$	$\sum X_2$		$\sum (X_2 - \bar{x}_2)^2$
$= 336$			$= 450$		$= 1410$

$$S^2 = \frac{1}{n_1 + n_2 - 2}\left[\sum (x_1 - \bar{x}_1)^2 + \sum (x_2 - \bar{x}_2)^2\right]$$

$$= \frac{1}{12 + 15 - 2}[380 + 1410]$$

$$= \frac{1}{25}[1790] = 71.6$$

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{S^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$= \frac{28 - 30}{\sqrt{71.6\left(\frac{1}{12} + \frac{1}{15}\right)}}$$

$$= \frac{-2}{\sqrt{71.6(0.083 + 0.066)}} = \frac{-2}{\sqrt{10.66}} = \frac{-2}{3.26} = -0.61$$

$$|t| = 0.61$$

d.f. $n_1 + n_2 - 2 = 25$ and table value of t for 25 d.f is 2.060 for 5% level of significance.

Since calculated value is less than table value at 5% level of significance, we Accept

the hypothesis.

Hence, we conclude that the two diets do not differ significantly as regards their effect on increase in weight.

Problem 112. *Samples of two types of electric light bulbs were tested for length of life and following data were obtained: Sample No.*

	Type - I	*Type - II*
$samplesize$	$n_1 = 8$	$n_2 = 7$
$SampleMeans$	$\bar{x}_1 = 1234hrs \bar{x}_2 = 1036hrs$	
$SampleS.D.$	$s_1 = 36hrs$	$s_2 = 40hrs$

Is the difference in the means sufficient to warrant that type I is superior to type II regarding length of life?

Solution: Setup Null and alternative hypothesis H_0 : There is no significance of differecne between two types of electric bulbs.

$$\bar{x}_1 = 1234; \bar{x}_2 = 1036; s_1 = 36; s_2 = 40$$

$$S^2 = \frac{1}{n_1 + n_2 - 2}\left[n_1 {s_1}^2 + n_2 {s_2}^2\right]$$

$$= \frac{1}{8 + 7 - 2}\left[8(36)^2 + 7(40)^2\right]$$

$$= \frac{1}{13}[8 \times 1296 + 7 \times 1600]$$

$$= \frac{1}{13}[10368 + 11200]$$

$$= \frac{1}{13}[21568]$$

$$= \frac{1}{13}[21568]$$

$$= 1659.08$$

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{S^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$= \frac{1234 - 1036}{\sqrt{1659.08\left(\frac{1}{8} + \frac{1}{7}\right)}}$$

$$= \frac{198}{\sqrt{1659.08 \times 0.2679}} = 9.37$$

$$|t| = 9.37$$

Since calculated value is greater than table value at 5% level of significance with $n_1 + n_2 - 2$ degrees of freedom, we Reject the hypothesis. Hence, we conclude that two types of bulbs differ significantly and type I is definitely superior to type II.

Problem 113. *A group of boys and girls were given an intelligence test. The mean score, S.D. score and numbers in each group are as follows.*

	Boys	*Girls*
Mean	*74*	*70.*
SD	*8*	*10*
n	*12*	*10*

Is the difference between the means of the two groups significant at 5% *level of significance* $t_{.05} = 2.086$. *for 20 d.f.*

Solution:(Solve it yourself !)
Hint steps: $s^2 = 88.4$, $s = 9.402$, $t = 0.994$

Problem 114. *Two types of batteries are tested for their length of life and the following results were obtained.*
Battery A : $n_1 = 10, \bar{x}_1 = 500$ *hrs,*$\sigma_1^2 = 100$
Battery B : $n_2 = 10, \bar{x}_2 = 560\text{hrs}, \sigma_2' = 121$
Compute Student's t *and test whether there is a significant difference in the two means.*

Solution: (Solve it yourself !)
Hint steps: $s^2 = 122.78$, $s = 11.0805$, $t = 12.11$, reject

Problem 115. *From a random sample of 10 pigs fed on diet* A*, the increases in weight in a certain period were* $10, 6, 16, 17, 13, 12, 8, 14, 15, 9$ *lbs. For another random sample of 12 pigs fed on diet* B*, the increases in the same period were* $7, 13, 22, 15, 12, 14, 18, 8, 21, 23, 10, 17\text{lbs}$. *Test whether diets* A *and* B *differ significantly as regards their effect on increases in weight?*

Solution:(Solve it yourself !)
Hint steps: $\bar{x} = 12$, $\bar{y} = 15$,
$s^2 = 21.1$, $s = 4.65$, $t = 1.6$

Problem 116. *Two horses A and B were tested according to the time (In Seconds) to run a particular race with the following results :*

Horse A	28	30	32	33	33	29	34
Horse B	29	30	30	24	27	29	–

Test whether you can discriminate between the two horses.

Solution: (Solve it yourself !)
Hint steps: $\bar{x} = 31.3$, $\bar{y} = 28.2$, $s = 2.3016$, $t = 2.42$

4.6 Testing of hypothesis: Chi-square test

In an experiment, the value of χ^2 gives a measure of the correspondence between theoretical and observed frequencies. Let $O_i(i = 1, 2, 3 \ldots n)$ and $E_i(i = 1, 2, 3 \ldots n)$ be the set of observed frequencies and expected frequencies respectively, then the Chi-square distribution is defined as

$$\chi^2 = \frac{(O_1 - E_i)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} + \frac{(O_3 - E_3)^2}{E_3} + \ldots\ldots + \frac{(O_n - E_n)^2}{E_n}$$

$$\Rightarrow \boxed{\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}}$$

where $\sum O_i = \sum E_i = N$(total frequency)

Chi – square test as a test of goodness of fit: χ^2 Test helps us to test the goodness of fit of the distributions such as Binomial, Poisson and Normal distributions. If the calculated value of χ^2 is less than the table value of χ^2 at a specified level of significance, the hypothesis is accepted. Otherwise the hypothesis is rejected.

Conditions for applying chi square test

(i) No theoretical(Expected) frequency should be smaller than 5. If any theoretical cell frequency is less than 5, then for the application of χ^2 test, the difficulty is overcome by grouping two or more classes together before calculating $(O_i - E_i)$. It is important to remember that the number of degrees of freedom is determined with the number of classes after regrouping.

(ii) $\sum O_i = \sum E_i = N$(total frequency)

Problem 117. *In an experiment of pea breeding, the following frequencies of seed were obtained.*

Round-yellow	*Wrinkled Yellow*	*Round Green*	*Wrinkled Green*	*Total*
315	*101*	*108*	*32*	*556*

Can you say that the experiment is in agreement with the theory which predicts proportion of frequencies, 9 : 3 : 3 : 1 ? Given that $\chi^2_{0.05} = 7.815$ *for 3 d.f. [VTU July 2013]*

Solution: Let H_0 : The experimental result support the theory i.e., there is no significant difference between the observed and theoretical frequency. Under H_0, the corresponding frequencies can be calculated as

$$\frac{9}{16} \times 556 = 312.75 \approx 313,$$
$$\frac{3}{16} \times 556 = 104.25 \approx 104,$$
$$\frac{3}{16} \times 556 = 104.25 \approx 104,$$
$$\frac{1}{16} \times 556 = 34.75 \approx 35$$

Hence the table of observed and theoretical frequencies is,

O_i	315	101	108	32	Sum=556
E_i	313	104	104	35	Sum=556

$$\begin{aligned}\chi^2 &= \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \\ &= \frac{(315-313)^2}{313} + \frac{(101-104)^2}{104} \\ &\quad + \frac{(108-104)^2}{104} + \frac{(32-35)^2}{35} \\ &= \frac{4}{313} + \frac{9}{104} + \frac{16}{104} + \frac{9}{35} \\ &= 0.51\end{aligned}$$

Also $d.f. = \nu = 4 - 1 = 3$

For $\nu = 3$, we have $chi^2_{0.05} = 7.815$ (by the table)

The calculated value of χ^2 is much less than $\chi^2_{0.05}$ so we accept H_0 and conclude that there is a very high degree of agreement between theory and practical.

Problem 118. *The following table gives the number of aircraft accidents that occured during the various days of the week. FInd whether the accident are uniformly disributed over the week.* ($\chi^2_{0.05} = 9.41$ *for 4 d.f.)*

Day	*Sun*	*Mon*	*Tue*	*Wed*	*Thu*	*Fri*	*Sat*	*Total*
No. Of Accidents	*14*	*16*	*8*	*12*	*11*	*9*	*14*	*84*

Given: The values of chi-square significant at $5, 6, 7$ *d.f. are respectively 11.07,12.59 14.07 at the* 5% *level of significance.* *[VTU June 2019, Dec 2012, June 2010]*

Solution: Null Hypothesis, H_0: The accidents are uniformly distributed over the week.

Under this H_0, the expected frequencies of the accidents on each of these days = $\frac{84}{7} = 12$

Expected frequencies of the accidents are given below :

O_i :	14	16	8	12	11	9	14	Sum= 84
E_i :	12	12	12	12	12	12	12	Sum= 84

$$\chi^2 = \sum_{i=1}^{n} \frac{(0_i - E_1)^2}{E_i}$$
$$= \frac{(14-12)^2}{12} + \frac{(16-12)^2}{12} + \frac{(8-12)^2}{12} + \frac{(12-12)^2}{12} + \frac{(11-12)^2}{12} + \frac{(9-12)^2}{12} + \frac{(14-12)^2}{12}$$
$$= \frac{1}{12}[4+16+16+0+1+9+4]$$
$$= \frac{50}{12} = 4.17$$

The number of degrees of freedom $= 7 - 1 = 6$

The tabulated $\chi^2_{0.05}$ for $6\text{d.f.} = 12.59$

since the calculated χ^2 is much less than the tabulated value, we accept the null hypothesis.

Hence, the accidents are uniformly distributed over the week.

Problem 119. *Fit a poisson distribution for the following data and test the goodness of fit in 5% level of significance. Given that* $\chi^2_{0.05} = 7.815$ *for 3 d.f*

x	*0*	*1*	*2*	*3*	*4*
frequency	*122*	*60*	*15*	*2*	*1*

[VTU: DEC/JAN 16]

Solution: Here $N = \Sigma f = 200$

Mean, $\mu = \bar{x} = \frac{\sum fx}{\Sigma f} = \frac{100}{200} = 0.5$

In Poisson distribution $\lambda = 0.5$, $P(x) = \frac{0.5^x e^{-0.5}}{x!}$

The Poisson distribution and the theoretical frequencies are calculated using

$$Np(x) = 200 \times \frac{\lambda^x e^{-\lambda}}{x!}$$
$$= 121.3 \times \frac{\lambda^x}{x!} \quad (\because 200e^{-0.5} = 121.3)$$

Hence theoretical frequencies are

$$121.3 \times \frac{(0.5)^0}{0!} = 121, \quad 121.3 \times \frac{(0.5)^1}{1!} = 61,$$
$$121.3 \times \frac{(0.5)^2}{2!} = 15, \quad 121.3 \times \frac{(0.5)^3}{3!} = 3,$$
$$121.3 \times \frac{(0.5)^4}{4!} = 0$$

Therefore new table with observed and expected(theoretical) frequencies is

O_i	122	60	15	2+1=3	Sum= 200
E_i	121	61	15	3+0=3	Sum= 200

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$
$$= \frac{(122-121)^2}{121} + \frac{(60-61)^2}{61} + \frac{(15-15)^2}{15} + \frac{(3-3)^2}{3}$$
$$= 0.025 < \chi^2_{0.05} = 7.815$$

Therefore the fitness is considered good.

$\therefore$ The hypothesis that the fitness is good can be accepted.

Problem 120. *Fit a binomial distribution for the data*

No. of Heads	*0*	*1*	*2*	*3*	*4*
Frequency	*122*	*60*	*15*	*2*	*1*

and also test the goodness of fit given that $\chi^2_{0.05} = 7.815$ *for 3 d.f. (VTU Model 2020)*

Solution: Here $N = \Sigma f = 200$

Mean, $\mu = \bar{x} = \frac{\sum fx}{\Sigma f} = \frac{100}{200} = 0.5$

i.e. $np = 0.5 \Rightarrow p = \frac{0.5}{n} = \frac{0.5}{4} = 0.125$

$\therefore q = 1 - p = 0.875$

Given Observed frequencies are 122,60,15,2,1

The expected frequencies are calculated by using $Np(x)$ where $N = 200$ and $p(x) = {}^nC_x p^x q^{n-x}, \ x = 0, 1, 2, 3, 4$

$NP(0) = (200)\ {}^4C_0 p^0 q^{4-0} = 117.23 \approx 117$

$NP(1) = (200)\ {}^4C_1 p^1 q^{4-1} = 66.99 \approx 67$

$NP(2) = (200)\ {}^4C_2 p^2 q^{4-2} = 14.36 \approx 14$

$NP(3) = (200)\ {}^4C_3 p^3 q^{4-3} = 1.36 \approx 2$ (Adjusted to get $N = 200$)

$NP(4) = (200)\ {}^4C_4 p^4 q^{4-4} = 0.05 \approx 0$

Therefore new table with observed and expected(theoretical) frequencies is

O_i	122	60	15	2+1=3	Sum= 200
E_i	117	67	14	2+0 =2	Sum= 200

$$\therefore \chi^2 = \sum_i \left[\frac{(0_i - E_i)^2}{E_i}\right]$$

$$= \frac{(122-117)^2}{117} + + \frac{(60-67)^2}{67} + \frac{(15-14)^2}{14} + \frac{(3-2)^2}{2}$$

$$= \frac{25}{117} + \frac{49}{67} + \frac{1}{14} + \frac{1}{2}$$

$$= 1.516 < \chi^2_{0.05} = 7.815 \text{(for 3 d.f.)}$$

Hence the fitness is good.

Problem 121. *A set of 5 similar coins tossed 320 times gives following table:*

No. of heads	*0*	*1*	*2*	*3*	*4*	*5*
Frequency	*6*	*27*	*72*	*112*	*71*	*32*

Test the hypothesis that data follows binomial distribution distribution. ($\nu = 5, \chi^2_{0.05} = 11.07$)

[VTU:JUNE/JULY-15, Jan 2015, July 2013, 2004]

Solution: Let H_0 : Data follows a binomial distribution.

Then H_1: data does not follow binomial distribution.

Given that level of significance is, $\alpha = 5\%$, with $n = 5$,

degrees of freedom is $\nu = 5$.

Here $N = \sum f = 320$

Let us compute the various expected frequencies.

Here for one toss of a coin $p = p(Head) = 0.5$
and $q = 1 - p = 0.5$
As the data is set to be following binomial distribution, the expected frequencies are given by $Np(x)$ for $x = 0, 1, 2, 3, 4, 5$ where $p(x) = {}^nC_x p^x q^{n-x} = {}^nC_x (0.5)^x (0.5)^{n-x} = {}^nC_x (0.5)^n$
Theoretical frequencies are $NP(0) = 320 \times {}^5C_0 p^0 q^{n-0} = 320 \times \left(\frac{1}{2}\right)^5 = 10$
$NP(1) = 320 \times {}^5C_1 p^1 q^{5-1} = 50$
$NP(2) = 320 \times {}^5C_2 p^2 q^3 = 100$
$NP(3) = 320 \times {}^5C_3 p^3 q^2 = 100$
$NP(4) = 320 \times {}^5C_4 p^4 q = 50$
$NP(5) = 320 \times {}^5C_5 p^5 q^0 = 10$
Thus, expected frequencies E_i are respectively $10, 50, 100, 100, 50, 10$.
Table of Observed and Expected frequencies is

O_i	6	27	72	112	71	32
E_i	10	50	100	100	50	10

Consider the test criterion given by

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$$
$$= \left(\frac{(6-10)^2}{10}\right) + \left(\frac{(27-50)^2}{50}\right) + \left(\frac{(72-100)^2}{100}\right)$$
$$+ \left(\frac{(112-100)^2}{100}\right) + \left(\frac{(71-50)^2}{50}\right) + \left(\frac{(32-10)^2}{10}\right)$$
$$= 78.68.$$

As the calculated value is very much higher than the tabulated value of $\chi^2_{0.05} = 11.07$ for $\nu = 5$ d.f., we reject the null hypothesis and accept the alternate hypothesis that data does not follow the binomial distribution.

Problem 122. *Four coins are tossed 100 times & following Results were obtained.*

No. of heads	*0*	*1*	*2*	*3*	*4*
Frequency	*5*	*29*	*36*	*25*	*5*

Fit a binomial distribution for the data and test the goodness of fit. ($\nu = 4, \chi^2_{0.05} = 9.49$) *[VTU: Model 2020, Dec 2018]*

Solution : Given the 4 coins are tossed 100 times

The probability of getting head is $\mathbf{p = 0.5, q = 0.5}$

The probability mass function of a binomial distribution is $\boldsymbol{P(x) = nC_x p^x q^{n-x} \quad x = 0, 1, 2, \ldots n}$

Expected (or theoretical) frequencies are given by $\boldsymbol{N \times P(x)}$ where $\boldsymbol{N = \Sigma f = 100}$ for $\boldsymbol{x = 0, 1, 2, \ldots n}$

$$\boldsymbol{NP(X = x) = N \times {}^4C_x (0.5)^x (0.5)^{4-x}}$$

$$\boldsymbol{NP(0) = N \times {}^4C_0 (0.5)^0 (0.5)^{4-0} = 100 \times 0.0625 = 6.25 \approx 6}$$

$$\boldsymbol{NP(1) = N \times {}^4C_1 (0.5)^1 (0.5)^{4-1} = (100)(0.25) = 25}$$

$$\boldsymbol{NP(2) = N \times {}^4C_2 (0.5)^2 (0.5)^{4-2} = (100)0.375 = 37.5 \approx 38}$$

$$\boldsymbol{NP(3) = N \times {}^4C_3 (0.5)^3 (0.5)^{4-3} = (100)0.25 = 25}$$

$$\boldsymbol{NP(4) = N \times {}^4C_4 (0.5)^3 (0.5)^{4-4} = (100)0.0625 = 6.25 \approx 6}$$

Hence the table of observed and theoretical frequencies is,

$\boldsymbol{O_i}$	5	29	36	25	5	$\boldsymbol{Sum = 100}$
$\boldsymbol{E_i}$	6	25	38	25	6	$\boldsymbol{Sum = 100}$

$$\therefore \chi^2 = \sum \left[\frac{(O_i - E_i)^2}{E_i} \right]$$

$$\Rightarrow \chi^2 = \frac{1}{6} + \frac{16}{25} + \frac{4}{38} + 0 + \frac{1}{6}$$

$$= 1.0786$$

$\chi^2 = 1.0786 < \chi^2_{0.05} = 9.49$ for 4 d.f.

Hence we accept $\boldsymbol{H_0}$ and conclude that the fitness is good.

Problem 123. *A die was thrown 60 times and the following frequency distribution was observed:*

Faces	*1*	*2*	*3*	*4*	*5*	*6*
Frequency	*15*	*6*	*4*	*7*	*11*	*17*

Test whether the die is unbiased at $\mathbf{5\%}$ *significance level.*

Solution: The frequencies in the given data are the observed frequencies.

Assuming that dice is unbiased, the expected number of frequencies for the numbers $\mathbf{1, 2, 3, 4, 5, 6}$ to appear on the face is $\frac{60}{6} = 10$ each.

Now the data is as follows:

x	1	2	3	4	5	6
O_i	15	6	4	7	11	17
E_i	10	10	10	10	10	10

$$\chi^2 = \sum_i \left[\frac{(O_i - E_i)^2}{E_i}\right]$$
$$= \frac{(15-10)^2}{10} + \frac{(15-6)^2}{10} + \frac{(15-4)^2}{10} + \frac{(15-7)^2}{10} + \frac{(15-11)^2}{10} + \frac{(15-17)^2}{10}$$
$$= \frac{1}{10}[25 + 81 + 121 + 64 + 16 + 4] = \frac{311}{10}$$
$$= 31.10$$

4.7 F-test

F test(Fisher's test) is a statistical test that is used in hypothesis testing to check whether the variances of two populations or two samples are equal or not or to test if the two samples have come from same population. This test uses the F statistic to compare two variances by dividing them.

Let $x_1, x_2, \ldots x_{n_1}$ and $y_1, y_2, \ldots y_{n_2}$ be the values of two independent random samples drawn from the normal populations σ^2 having equal variances.

Let $\bar{x}_1$ and $\bar{x}_2$ be the sample means and

$$s_1^2 = \frac{1}{n_1 - 1}\sum_1^{n_1}(x_i - \bar{x})^2$$
$$s_2^2 = \frac{1}{n_2 - 1}\sum_1^{n_2}(y_i - \bar{y})^2$$

be the sample variances.

Then we define F by the relation

$$F = \frac{s_1^2}{s_2^2}$$
$$(s_1^2 > s_2^2)$$

This gives F-distribution (also known as variance ratio distribution) with $\nu_1 = n_1 - 1$ and $\nu_2 = n - 2$ degrees of freedom. The larger of the variances is placed in the numerator.

Problem 124. *The I.Q.'s of 25 students from one college showed a variance of 16 and those of an equal number from the other college had a variance of 8 . Discuss whether there is any significant difference in variability of intelligence.*

Solution: $\sigma_1{}^2 = 16, \sigma_2{}^2 = 8$

$$F = \frac{\sigma_1{}^2}{\sigma_2{}^2} = \frac{16}{8} = 2$$

Tabulated value of F at 5% level of significance $= 1.98$
Calculated value of $F = 2$ Tabulated value of $F(1.98)$
Hence, variability of intelligence is just significant at 5% level of significant.
Tabulated value of F at 1% level of significance $= 2.62$
Calculated value $F = 2 <$ Tabulated value of $F(2.62)$
Hence, variability of intelligence is not significant at 1% level of significance.

Problem 125. *Two samples of sizes 9 and 8 give the sum of squares of deviations from their respective means equal to 160 inches ^{2} and 91 inches ^{2} respectively. Can these be regarded as drawn from the same normal population?*

Solution: Let H_0: There is no significant difference between the variances of two populations.
We have $\Sigma(x - \bar{x})^2 = 160$ and $\Sigma(y - \bar{y})^2 = 91$ $\therefore$

$$s_1^2 = \frac{160}{8} = 20$$

and

$$s_2^2 = \frac{91}{7} = 13.$$

Hence

$$F = \frac{s_1^2}{s_2^2} = \frac{20}{13} = 1.54 \text{ nearly.}$$

For $\nu_1 = n_1 - 1 = 8, \nu_2 = n_2 - 1 = 7$, we have $F_{0.05} = 3.73$
Since the calculated value of $F < F_{0.05}$, we accept H_0.
i.e. the population variances are not significantly different.
Thus the two samples can be regarded as drawn from two normal populations with the same variance.

Problem 126. *Two independent samples of sizes 7 and 6 have the following values :*
Sample A: 28 30 32 33 33 29 34
Sample B : 29 30 30 24 27 29
Examine whether the samples have been drawn from normal populations having the

same variance?[Given that the values of F *at* 5% *level for* $(6,5)$ *d.f. is 4.95 and for* $(5,6)$ *d.f. is 4.391]*

Solution: H_0: The samples have been drawn from normal populations having the same variance.

$$\bar{x} = \frac{219}{7} = 31.285 \text{ and } \bar{y} = \frac{169}{6} = 28.166.$$

Then

$$\begin{aligned} S_1^2 =& \frac{1}{n_1 - 1} \sum (x_i - \bar{x})^2 \\ =& \frac{1}{6} \left[(28 - 31.285)^2 + (30 - 31.285)^2 \right. \\ & + (32 - 31.285)^2 + (33 - 31.285)^2 \\ & + (33 - 31.285)^2 + (29 - 31.285)^2 \\ & \left. + (34 - 31.285)^2 \right] \\ =& \frac{1}{6} [10.791 + 1.651 + 0.511 + 2.941 \\ & + 2.941 + 5.221 + 7.371] = 5.238 \end{aligned}$$

and

$$\begin{aligned} S_2^2 =& \frac{1}{n_2 - 1} \sum (y_i - \bar{y})^2 \\ =& \frac{1}{5} \left[(29 - 28.166)^2 + (30 - 28.166)^2 \right. \\ & + (30 - 28.166)^2 + (24 - 28.166)^2 \\ & \left. + (27 - 28.166)^2 + (29 - 28.166)^2 \right] \\ =& \frac{1}{5} [0.695 + 3.364 + 3.364 + 17.355 \\ & + 1.359 + 0.695] = 5.366. \end{aligned}$$

Therefore, the test statistics is given by

$$F = \frac{S_2^2}{S_1^2} = \frac{5.366}{5.238} = 1.025.$$

Further, since numbers of degree of freedom are 6 and 5 , we have from table, $F_{0.05}(6,5) = 4.95$. Since the calculated value is less than the tabular value, we accept H_0.

i.e. The samples have been drawn from normal populations having the same variance.

Problem 127. *The nicotine content (in* $\mathbf{mg}$ *) of two samples of tobacco were found to be as follows :*

Sample A: 24 27 26 21 25

Sample B: 27 30 28 31 22 36

Can it be said that the two samples came from the same population?

Solution: H_0: The samples have been drawn from normal populations having the same variance.

Suppose that $\bar{x}$ be the sample mean for the sample B and $\bar{y}$ be the sample mean of the sample A. Then

$$\bar{x} = \frac{174}{6} = 29 \text{ and } \bar{y} = \frac{123}{5} = 24.6$$

$$\begin{aligned} S_1^2 &= \frac{1}{n_1 - 1} \sum (x_1 - \bar{x})^2 \\ &= \frac{1}{5} \left[(27-29)^2 + (30-29)^2 + (28-29)^2 \right. \\ &\quad \left. + (31-29)^2 + (22-29)^2 + (36-29)^2 \right] \\ &= \frac{1}{5}[4 + 1 + 1 + 4 + 49 + 49] = 21.6, \end{aligned}$$

$$\begin{aligned} S_2^2 &= \frac{1}{n_2 - 1} \sum (y_1 - \bar{y})^2 \\ &= \frac{1}{4} \left[(24-24.6)^2 + (27-24.6)^2 + (26-24.6)^2 \right. \\ &\quad \left. + (21-24.6)^2 + (25-24.6)^2 \right] \\ &= \frac{1}{4}[0.36 + 5.76 + 1.96 + 12.96 + 0.16] = 5.3. \end{aligned}$$

Therefore, the statistics for F-test is

$$F = \frac{S_1^2}{S_2^2} = \frac{21.6}{5.3} = 4.08.$$

But tabular value of $F(5, 4) = 6.26$.

Since the calculated value is less than the tabular value, we accept H_0.

i.e. The samples have been drawn from normal populations having the same variance

Question Bank:Module 4

1. An electrical firm manufactures light bulbs that have a length of life that is approximately normally distributed, with mean equal to 800 hours and a standard

deviation of 40 hours. Find the probability that a random sample of 16 bulbs will have an average life of less than 775 hours. Ans: 0.0062

2. A random sample $n = 100$ is taken from an infinte Population with mean $\mu = 75$ And variance = 256. Based on central limit theorem With what probability can we Assert the value We obtain for X will fall between 67 and 83? Hint : Ans: $A(-0.3125 to 0.3125) = 0.376$

3. An unknown distribution has a mean of 90 and a standard deviation of 15 . A sample size of 80 is drawn randomly from the population. Find the probability that the sum of the 80 values (or the total of the 80 values) is more than 7400 . Ans: 0.0681

4. An unknown distribution has a mean of 90 and a standard deviation of 15 . Samples of size $n = 25$ are drawn randomly from the population. Find the probability that the sample mean is between 85 and 92 . Ans:0.6997

5. State Central limit theorem. Use the theorem to evaluate $P[50 < \bar{X} < 56]$ where $\bar{X}$ represents the mean of a random sample of size 100 from an infinite population with mean $\mu = 53$ and variance $\sigma^2 = 400$ (VTU Model 2023) Ans:0.8664

6. Certain tubes manufactured by a company have mean life time of 800 hours and S.D of 60 hours. Find the probability that a random sample of 16 tubes taken from the group will have a mean life time (a) between 790 hours and 810 hours. (b) less than 785 hours. (c) more than 820 hours. (d) between 770 hours and 830 hours. Ans:0.4972, 0.1587, 0.0918, 0.9544

7. A random sample of size 64 is taken from an infinite population having mean 112 and variance 144. Using central limit theorem, find the probability of getting the sample mean $\bar{X}$ greater than 114.5 (VTU Model 2023) Ans:0.0489

8. In a recent study reported on the Flurry Blog, the mean age of tablet users is 34 years. Suppose the standard deviation is 15 years. Take a sample of size $n = 100$. Using central limit theorem, find the probability that the sample mean age is more than 30 years. (VTU Model 2023) Ans:0.9961

9. A random sample of 10 boys had the following IQ :
   $\mathbf{70, 120, 110, 101, 88, 83, 95, 98, 107, 100.}$ Do these data support the assumption of a population mean $\mathbf{1Q}$ of $\mathbf{100?}$ (at $\mathbf{5\%}$ level of significance? [VTU Model 2020,2006] hint Ans: $\boldsymbol{t = 0.6203 < 2.262}$

10. Ten individuals are chosen at random from the population and their heights are found to be inches $\mathbf{63, 63, 64, 65, 66, 69, 69, 70, 70, 71}$. Discuss the suggestion that the mean height in the universe is $\mathbf{65}$ inches, given that for $\mathbf{9}$ degree of freedom the value of student's 't' at $\mathbf{0.05}$ level of significance is $\mathbf{2.262}$ hint Ans: $\boldsymbol{|t| = 2.02 < t_{0.05} = 2.262}$

11. Nine items have values $\mathbf{45, 47, 50, 52, 48, 47, 49, 53, 51}$. Does the mean of these differ significantly from assumed mean of $\mathbf{47.5?}(\boldsymbol{\nu = 8, t_{0.05} = 2.31})$ [VTU:JUNE/JULY-15, July 2013, 2010] hint Ans: $\boldsymbol{|t| = 1.84 < t_{0.05} = 2.31}$

12. A machinist is making engine parts with axle diameter of $\mathbf{0.7}$ inch. A random sample of $\mathbf{10}$ parts shows mean diameter of $\mathbf{0.742}$ inch with a standard deviation of $\mathbf{0.04}$ inch. On the basis of this sample, would you say that the work is inferior? Ans: $\boldsymbol{t = 3.32 > t_{0.05}}$.

13. certain stimulus administered to each of 12 patients resulted in the following increase of blood pressure : $\mathbf{5, 2, 8, -1, 3, 0, -2, 1, 5, 0, 4, 6}$. Can it be concluded that the stimulus will in general be accompanied by an increase in blood pressure? [VTU Model **2020**, June **2019**, Jan **2018**, Dec **2012**, Dec **2010, 2007**] Ans: $\boldsymbol{t = 2.924}$

14. A group of boys and girls were given an intelligence test. The mean score, S.D. score and numbers in each group are as follows.

	Boys	Girls
Mean	74	70.
$\boldsymbol{SD}$	8	10
$\boldsymbol{n}$	12	10

   Is the difference between the means of the two groups significant at $\mathbf{5\%}$ level of significance $\boldsymbol{t_{.05} = 2.086.}$ for 20 d.f. Ans:$\boldsymbol{t = 0.994}$

15. Two types of batteries are tested for their length of life and the following results were obtained.
   Battery $\boldsymbol{A : n_1 = 10, \bar{x}_1 = 500}$ hrs,$\boldsymbol{\sigma_1^2 = 100}$

Battery $B : n_2 = 10, \bar{x}_2 = 560\text{hrs}, \sigma_2' = 121$
Compute Student's t and test whether there is a significant difference in the two means. Ans:$t = 12.11$

16. From a random sample of 10 pigs fed on diet A, the increases in weight in a certain period were 10, $6, 16, 17, 13, 12, 8, 14, 15, 9$ lbs. For another random sample of 12 pigs fed on diet B, the increases in the same period were $7, 13, 22, 15, 12, 14, 18, 8, 21, 23, 10, 17$lbs. Test whether diets A and B differ significantly as regards their effect on increases in weight? Ans:$t = 1.6$

17. Two horses A and B were tested according to the time (In Seconds) to run a particular race with the following results :

Horse A	28	30	32	33	33	29	34
Horse B	29	30	30	24	27	29	–

Test whether you can discriminate between the two horses. Ans:$t = 2.42$

18. b) Suppose that $10, 12, 16, 19$ is a sample taken from a normal population with variance 6.25 . Find at 95% confidence interval for the population mean.

19. A sample of size 15 drawn from a normally distributed population has sample mean 35 and sample standard deviation 14. Construct a 95% confidence interval for the population mean. Ans : $(27.2, 42.8)$

20. A random sample of 12 students from a large university yields mean GPA 2.71 with sample standard deviation 0.51. Construct a 90% confidence interval for the mean GPA of all students at the university. Assume that the numerical population of GPAs from which the sample is taken has a normal distribution. Ans :$(2.45, 2.97)$.

21. For a sample of 10 cows that had recently given birth, the mean protein content in 40oz of milk was 3.4 g with a sample standard deviation of 0.4 g. Find a 95% confidence interval for the mean protein content of the milk. Ans: (3.19,3.61)

22. A simple random sample of 20 statistics students during a statistics exam gives an average pulse rate 74.4 with a standard deviation of 10. (a) Find 90%, 95%, 99% confidence intervals for the average pulse rate of all statistics students during an exam. Ans : (70.53,78.27), (69.72,79.08) and (68.00,80.80)

23. 100 packs were tested and the mean weight was 24.1 grams. Calculate a 99% confidence interval for the population mean μ. Ans: (23.69, 24.51)

24. In an experiment of pea breeding, the following frequencies of seed were obtained.

Round-yellow	Wrinkled Yellow	Round Green	Wrinkled Green	Total
315	101	108	32	556

Can you say that the experiment is in agreement with the theory which predicts proportion of frequencies,
9 : 3 : 3 : 1 ? Given that $\chi^2_{0.05} = 7.815$ for 3 d.f. Ans: $\chi^2 = 0.51$ ¡ $\chi^2_{0.05} = 7.815$

25. The following table gives the number of aircraft accidents that occured during the various days of the week. FInd whether the accident are uniformly distributed over the week. ($\chi^2_{0.05} = 9.41$ for 4 d.f.) Ans: $\chi^2_{0.05} = 4.17 < \chi^2_{0.05} = 12.59$

26. A die was thrown 60 times and the following frequency distribution was observed:

Faces	1	2	3	4	5	6
Frequency	15	6	4	7	11	17

Test whether the die is unbiased at 5% significance level. Ans: 31.10

27. Fit a poisson distribution for the following data and test the goodness of fit in 5% level of significance. Given that $\chi^2_{0.05} = 7.815$ for 3 d.f

x	0	1	2	3	4
frequency	122	60	15	2	1

Ans:$\chi^2 < \chi^2_{0.05} = 7.815$

28. Fit a binomial distribution for the data

No. of Heads	0	1	2	3	4
Frequency	122	60	15	2	1

and also test the goodness of fit given that $\chi^2_{0.05} = 7.815$ for 3 d.f. Ans: $\chi^2 = 1.516 < \chi^2_{0.05} = 7.815$

29. A set of 5 similar coins tossed 320 times gives following table:

No. of heads	0	1	2	3	4	5
Frequency	6	27	72	112	71	32

Test the hypothesis that data follows binomial distribution distribution. ($\nu = 5, \chi^2_{0.05} = 11.07$) Ans:$\chi^2 = 78.68 > \chi^2_{0.05} = 11.07$

30. Four coins are tossed 100 times & following Results were obtained.

No. of heads	0	1	2	3	4
Frequency	5	29	36	25	5

Fit a binomial distribution for the data and test the goodness of fit. ($\nu = 4, \chi^2_{0.05} = 9.49$) Ans: $\chi^2 = 1.0786 < \chi^2_{0.05} = 9.49$

31. The I.Q.'s of 25 students from one college showed a variance of 16 and those of an equal number from the other college had a variance of 8 . Discuss whether there is any significant difference in variability of intelligence. Use 1% and 5% levels of significance. Ans:$F = 1.54 < F_{0.05} = 3.73$

32. Two samples of sizes 9 and 8 give the sum of squares of deviations from their respective means equal to 160 inches ^{2} and 91 inches ^{2} respectively. Can these be regarded as drawn from the same normal population? Hint Ans: $F = 1.54 < F_{0.05} = 3.73$

33. Two independent samples of sizes 7 and 6 have the following values :
Sample A: 28 30 32 33 33 29 34
Sample B : 29 30 30 24 27 29
Examine whether the samples have been drawn from normal populations having the same variance? Ans:$F = 1.025$

34. In two groups of ten children each, the increase in weight due to different diets during the same period, were in pounds
$3, 7, 5, 6, 5, 4, 4, 5, 3, 6$
$8, 5, 7, 8, 3, 2, 7, 6, 5, 7.$
Is there a significant difference in their variability? Ans:$F = 2.4 < 3.2$

Module 5

Design of Experiments & ANOVA

Syllabus

Principles of experimentation in design, Analysis of completely randomized design, randomized block design. The ANOVA Technique, Basic Principle of ANOVA, One-way ANOVA Completely Randomized Design.
Two-way ANOVA, Latin-square Design, and Analysis of Co-Variance.

5.1 Design and Analysis of experiments

Planning an experiment to gather appropriate data and drawing meaningful inference from that data with respect to any problem under investigation is known as design and analysis of experiments. Before conducting an experiment, an **experimental unit** is to be defined. For example, a leaf, a tree or a collection of adjacent trees may be an experimental unit. An experimental unit is also sometimes referred as **plot**. A collection of plots is termed a **block**. Observations made on experimental units differ considerably. These differences (variations) are partly produced by the manipulation of certain variables of interest generally called **treatments**, built-in and intentionally changed in the experiment to study their influences. For instance, in a fertilizer trial, the different types of fertilizers are treatments. However, not all differences come from these treatments. Besides the variations produced in the observations due to these known sources, the variations are also produced by a large number of unknown sources (Sources of Variation), such as uncontrolled genetic variations in plants.

5.1.1 Three basic principles of experimental designs :

Almost all experiments involve the three basic principles, viz., randomization, replication and local control.

The **Principle of Replication** suggests repeating experiments to get more reliable results. Instead of testing a treatment on just one subject, we apply it to many. For example, if comparing two types of rice, instead of growing each in only one part of a field, we grow them in several parts. This makes our conclusions more dependable. While this may introduce some computational challenges, it is essential for improving accuracy and getting more precise results. Replication helps in reducing the impact of random variations, making our findings more precise.

The **Principle of Randomization** is like a shield for experiments. It suggests making choices randomly to protect against sneaky outside factors. For example, when planting different rice types, instead of deciding where each one goes, we use randomness. This helps us avoid problems like uneven soil fertility. By using randomization, we ensure that any hidden factors are balanced out by chance, giving us more accurate results. It's like letting luck even things out, making our conclusions more trustworthy.

The **Principle of Local Control** is like a detective strategy for experiments. Imagine you are testing different rice types and want to rule out the impact of varying soil fertility. Here Here's what you do: you Divide the field into similar sections (homogenous blocks) and split each block into parts for each rice type. Randomly assign treatments within each block to spread out soil fertility differences. This way, you can precisely measure and remove the impact of soil fertility from your final results. In simple terms, local control helps you focus on the rice types by keeping other factors, like soil fertility, in check.

Example:

Consider a plant growth experiment:

Experimental Unit: Individual plants.

Treatments: Varying amounts of sunlight exposure.

Randomization: Randomly assign plants to different sunlight levels.

Replication: Repeat the experiment with multiple plants for each sunlight level.

Local Control: Ensure consistent soil conditions for all plants.

By applying these principles, we can draw conclusions about how different sunlight levels impact plant growth while minimizing the influence of other variables.

ANOVA (Analysis of Variance):

Analysis of Variance (ANOVA) is a statistical method used to test differences between two or more sample means. The name "Analysis of Variance" is used because inferences about means are made by analysing variance.
This technique is used when multiple sample cases or treatments are involved. Using this technique, one can draw inferences about whether the samples have been drawn from populations having the same mean.
In ANOVA, we have to make two estimates of population variance:
(i) one based on between samples variance(Cause Variance or Treatment variance) and
(ii) the other based on within samples variance(Chance Variance or Error variance or random variance or residual variance.).
Then the said two estimates of population variance are compared with F-test, given by

$$F = \frac{\text{Estimate of population variance based on between samples variance}}{\text{Estimate of population variance based on within samples variance}}$$

Example:
Consider a study comparing the effectiveness of three teaching methods:
Between Samples Variance: Examines how much the mean scores differ among the three teaching methods.
Within Samples Variance: Assesses the variability in scores within each teaching method group.
Using the F-test, we can determine if the observed differences in mean scores among teaching methods are statistically significant or merely due to chance.

5.2 Completely Randomized Design (CRD):

A completely randomized design (CRD) is one where the treatments are assigned completely at random to experimental units, so that each experimental unit has the same chance of receiving any particular treatment.
Example:
Consider a study evaluating the growth of plants with three different fertilizers. In a CRD, each plant is randomly assigned one of the three fertilizers. This ensures that any observed differences in plant growth are due to the fertilizers and not because of

any intentional patterns in how we assigned the fertilizers. We want to be sure that any changes in growth are connected to the specific fertilizers and not due to how we chose which plant gets which fertilizer.

One-way (or single factor) ANOVA or Completely Randomized Design (CRD) :

Under the one-way ANOVA, we consider only **one factor or property, or characteristic,** and then determine if there are differences within that factor. The experimental units are randomly assigned to different levels of the single factor or treatment. The analysis involves comparing the means of different treatment groups.
For example, we might want to know if three different groups of students have different mean marks. To see if there is a statistically significant difference in mean marks, we can conduct a one-way ANOVA.
The technique involves the following steps:
Consider k samples $x_1, x_2, x_3, \cdots, x_k$.
Let $x_{11}, x_{12}, x_{13}, \cdots, x_{1n_1}$ be the observations in the first sample.
Let $x_{21}, x_{22}, x_{23}, \cdots, x_{2n_2}$ be the observations in the second sample.
In this way let $x_{k1}, x_{k2}, x_{k3}, \cdots, x_{kn_k}$ be the observations in the kth sample.

Steps involved in one way ANOVA:

Step 1. Define the null hypothesis $H_0 : \mu_1 = \mu_2 = \mu_3 = \ldots = \mu_n$ for the level of significance.

Step 2. Let n_i be the number of items in ith sample and $N = \sum n_i$=total number of observations.

Step 3. Find the sum of observations in the i th sample,(T_i)
and find the Grand total, $T = \sum T_i = \sum x_{ij}$ = The sum of of all N observations where $i = 1, 2, 3, \ldots k$

Step 4. Find the correction factor, using

$$\text{CF} = \frac{T^2}{N}$$

Step 5. Find the squares of all observations and then find the total sum of squares

$$\text{TSS} = \sum_i \sum_j (x_{ij})^2 - \text{CF}$$

Step 6. Find the sum of the squares between the samples(treatments)

$$\text{SSB} = \sum_i \frac{T_i^2}{n_i} - \text{C.F.} \qquad i = 1, 2, 3, \ldots, k$$

where subscript i represents different categories.

Step 7. The sum of squares within the samples can be found out by subtracting SSB from TSS.

SSE= TSS-SSB

Step 8. The degrees of freedom for total sum of squares (TSS) is $(N-1)$. The degrees of freedom for SSB is $(k-1)$ and the degrees of freedom for SSE is $(N-k)$

Step 9. Mean sum of squares : The mean sum of squares for treatments is

$$S_1^2 = \frac{\text{SSB}}{\text{k}-1}$$

and the mean sum of squares for error is

$$S_2^2 = \frac{\text{SSE}}{\text{N}-\text{k}}$$

.

Step 10. ANOVA Table: The above sum of squares together with their respective degrees of freedom and mean sum of squares can be summarized in the ANOVA table in the following way.

Source of Variation	Sum of Squares	Degrees of freedom	Mean Sum of Square	F- Ratio(calculted)
Between Samples	SSB	$k-1$	$\text{S}_1^2 = \frac{\text{SSB}}{k-1}$	$\text{F} = \frac{S_1^2}{S_2^2}$
Within Samples(error)	SSE	$N-k$	$\text{S}_2^2 = \frac{\text{SSE}}{\text{N}-\text{k}}$	

Step 11. Calculation of F : Variance ratio F is the ratio between greater variance and smaller variance, thus

$$F = \frac{S_1^2}{S_2^2}$$

If variance within the treatment is more than the variance between the treatments, then numerator and denominator should be interchanged and degrees of freedom adjusted accordingly.

Step 12. Critical value of F or table value of F : The Critical value of F or table value of F is obtained from F distribution table for $(\mathbf{k-1,\ N-k})$ d.f at 5% level of significance.

Step 13. Inference: If calculated F value is less than table value of F, we accept our null hypothesis H_0 and say that there is no significant difference between treatments.
If calculated F value is greater than table value of F, we reject our H_0 and say that the difference between treatments is significant.

One Way ANOVA problems

Problem 128. *A test was given to five students taken at random from the fifth class of three schools of a town. The individual scores are*

School I	*9*	*7*	*6*	*5*	*8*
School II	*7*	*4*	*5*	*4*	*5*
School III	*6*	*5*	*6*	*7*	*6*

Carry out the analysis of variance.

Solution:
To carry out the analysis of variance, we form the following tables.

						Total	Squares
School I	9	7	6	5	8	35	1225
School II	7	4	5	4	5	25	625
School III	6	5	6	7	6	30	900
					Total	$T = 90$	2750

Table of Squares:

School I	81	49	36	25	64
School II	49	16	25	16	25
School III	36	25	36	49	36

Test Procedure: Null Hypothesis: $\mathbf{H_0 : \mu_1 = \mu_2 = \mu_3}$ i.e., There is no significant difference between the performance of schools.

Alternative Hypothesis $\mathbf{H_1 : \mu_1 \neq \mu_2 \neq \mu_3}$
Level of significance : Let $\boldsymbol{\alpha : 0.05}$
Test statistic Correct factor $\mathbf{(C.F)} = \frac{G^2}{N}$

$$= \frac{90^2}{15}$$

$$= \frac{8100}{15} = 540$$

Total sum of squares (TSS) $= \Sigma\Sigma {x_{ij}}^2 - \mathbf{C.F}$

$$= 568 - 540 = 28$$

Sum of squares between schools $= \frac{\sum T_i^2}{n_i} - \text{C.F}$

$$= \frac{2750}{5} - 540$$

$$= 550 - 540 = 10$$

Sum of squares due to error $\mathbf{(SSE)} = \text{TSS} - \text{SST}$

$$= 28 - 10 = 18$$

ANOVA Table

Sources of variation	d.f	S.S	M.S.S	F ratio
Between Schools	$3 - 1 = 2$	10	$\frac{10}{2} = 5.0$	$\frac{5}{1.5} = 3.33$
Error	12	18	$\frac{18}{12} = 1.5$	
Total	$15 - 1 = 14$			

Table Value: Table value of $\mathbf{F_e}$ for $\mathbf{(2, 12)}$ d.f at 5% level of significance is 3.8853
Inference: Since calculated $\boldsymbol{F_0}$ is less than table value of $\boldsymbol{F_e}$, we may accept our H0 and say that there is no significant difference between the performance of schools.

Problem 129. *Three processes A, B and C are tested to see whether their outputs are equivalent. The following observations of outputs are made:*

A	*10*	*12*	*13*	*11*	*10*	*14*	*15*	*13*
B	*9*	*11*	*10*	*12*	*13*			
C	*11*	*10*	*15*	*14*	*12*	*13*		

Carry out the analysis of variance and state your conclusion.

Solution: Solve yourself !

Hint: T=228, CF=2736, TSS=58, SSB=7,

SSE=51, $S_1^2 = 3.5, S_2^2 = 3.19$,

F=1.097¡F(2,16)=3.63, accept H_0

Problem 130. *Suppose in an industrial experiment that an engineer is interested in how the mean absorption of moisture in concrete varies among 5 different concrete aggregates. The samples are exposed to moisture for 48 hours. It is decided that 6 samples are to be tested for each aggregate, requiring a total of 30 samples to be tested. The data are recorded in the following Table:*

Aggregate:	1	2	3	4	5	
	551	595	639	417	563	
	457	580	615	449	631	
	450	508	511	517	522	
	731	583	573	438	613	
	499	633	648	415	656	
	632	517	677	555	679	
Total	3320	3416	3663	2791	3664	16,854

Solution: Solve yourself !

Hint:

$SST = 209,377, \quad SSA = 85,356$

$SSE = 209,377 - 85,356 = 124,021.$

$F = 4.30 > F_{0.05} = 2.76$, REJECT H_0

Problem 131. *As head of the department of a consumers research organization you have the responsibility of testing and comparing life times of 4 brands of electric bulbs.suppose you test the life time of 3 electric bulbs each of 4 brands,the data is given below,each entry representing the life time of an electric bulb,measured in hundreds of hours.*

A	*B*	*C*	*D*
20	25	24	23
19	23	20	20
21	21	22	20

Solution: Solve yourself !

Hint: T=258, CF=5547, TSS=39, SSB=15,

SSE=24, $F = 1.67 < f(3,8) = 4.07$, ACCEPT H_0

Problem 132. *Set up an analysis of variance table for the following per acre production data for three varieties of wheat, each grown on 4 plots and state if the variety differences are significant.*

Plot of land	*Per acre production data*		
	Variety of wheat		
	A	***B***	***C***
1	*6*	*5*	*5*
2	*7*	*5*	*4*
3	*3*	*3*	*3*
4	*8*	*7*	*4*

Solution: Solve yourself !

Hint: T=60, TSS=32, SSB=8, SSE=24, $F = 1.5 < F(2,9) = 4.26$, ACCEPT H_0

Problem 133. *The varieties of wheat A, B, C were sown in 4 plots each and the following yields in quintals per acre were obtained.*

A	*8*	*4*	*6*	*7*
B	*7*	*6*	*5*	*3*
C	*2*	*5*	*4*	*4*

Test the significance of difference between the yields of varieties, given that 5% *tabulated value of* $\mathbf{F}$ *for 2 and 9 degrees of freedom is 4.26 .*

Solution: Solve yourself !

Hint: T=61, CF=310.08, TSS=34.92, SSB=12.67, SSE=22.25,

$S_1^2 = 6.34, S_2^2 = 2.47, F = 2.57$

Problem 134. *Three different machines are used for a production. On the basis of the outputs, set up one - way ANOVA table and test whether the machines are equally effective.*

Outputs		
Machine I	*Machine II*	*Machine III*
10	*9*	*20*
15	*7*	*16*
11	*5*	*10*
10	*6*	*14*

Given that the value of F *at* 5% *level of significance for* $(2,9)$d. f *is 4.26*

Solution: Solve yourself !

Hint: CF=1704.08, TSS=284.92, SSB=162.17, SSE=122.75,

$S_1^2 = 81.085, S_2^2 = 13.63, F = 5.95 > F(2,9) = 4.26$, reject

Problem 135. *A manufacturing company has purchased three new machines of different makes and wishes to determine whether one of them is faster than the others in Producing a certain output. Five hourly production figures are observed at random from each other machine and the results are given below:*

Observation	A_1	A_2	A_3
1	*25*	*31*	*24*
2	*30*	*39*	*30*
3	*36*	*38*	*28*
4	*38*	*42*	*25*
5	*31*	*35*	*28*

Use ANOVA and determine whether the machines are significantly different in their mean speed.

Solution: Solve yourself ! Hint : Since values are slight larger, you can use coding method. Previous method can also be used.

Problem 136. *Three different kinds of food are tested on three groups of rats for 5 weeks. The objective is to check the difference in mean weight (in grams) of the rats per week. Apply one-way ANOVA using a 0.05 significance level to the following data:*

Food 1	8	12	19	8	6	11
Food 2	4	5	4	6	9	7
Food 3	11	8	7	13	7	9

Solution: Solve yourself !

Hint: Ans: N=18, T=154, cf=1316, TSS=230, SSR=75,

SSE=155, $S_1^2 = 37.5, S_2^2 = 9.1$, F=4.121, reject

5.3 TWO-WAY ANOVA or Randomized Block Design

Two-way ANOVA technique is used when the data are classified on the basis of two factors. In this, The experimental units are grouped into blocks based on one factor, and treatments are then randomly assigned within each block. **For example,**

- The agricultural output may be classified on the basis of different varieties of seeds and also on the basis of different varieties of fertilizers used.
- A business firm may have its sales data classified on the basis of different salesmen and also on the basis of sales in different regions.
- In a factory, the various units of a product produced during a certain period may be classified on the basis of different varieties of machines used and also on the basis of different grades of labour.

Such a two-way design may have repeated measurements of each factor or may not have repeated values. We shall now explain the two-way ANOVA technique in the context of both the said designs with the help of examples.

5.4 Steps involved in Two- way ANOVA(when repeated values are not there):

The various steps involved are as follows:

Step 1. Define the null hypothesis
H_0 : There is no significant difference between the means on the basis of row factors and there is no significant difference between the means on the basis of column factors.

Step 2. Let n_i be the number of items in ith sample and $N = \sum n_i$=total number of observations.

Step 3. Find the sum of observations in the i th sample,(T_i)
and find the Grand total, $T = \sum T_i = \sum x_{ij}$ = The sum of of all N observations where $i = 1, 2, 3, \ldots r$, where r is the number of rows.

Step 4. Find the correction factor, using

$$\text{CF} = \frac{T^2}{N}$$

Step 5. Find the squares of all observations and then find the total sum of squares

$$\text{TSS} = \sum_i \sum_j (x_{ij})^2 - \text{CF}$$

Step 6. Take the total of different rows, T_i and then obtain the square of each row total (i.e. T_i^2) and divide such squared values of each row by the number of items in the concerning row(i.e. n_i) and take the total of the result thus obtained. Finally, subtract the correction factor from this total to obtain the sum of squares of deviations for variance between rows (SSR).

$$\text{SSR} = \sum_i \frac{T_i^2}{n_i} - \text{C.F.} \qquad i = 1, 2, 3, \ldots r$$

, where r is the number of rows. (where subscript i represents ith row).

Step 7. Take the total of different columns, T_j and then obtain the square of each column total (i.e. T_j^2) and divide such squared values of each column by the number of items in the concerning column(i.e. n_j) and take the total of the result thus obtained. Finally, subtract the correction factor from this total to obtain the sum of squares of deviations for variance between columns (SSC).

$$\text{SSC} = \sum_j \frac{T_j^2}{n_j} - \text{C.F.} \qquad j = 1, 2, 3, \ldots c,$$

where c is the number of columns.(where subscript j represents jth column).

Step 8. The Sum of squares of deviations for residual or error can be found out by subtracting sum of SSB and SSC from TSS.

SSE= TSS-(SSB+SSC)

Step 9. Degrees of freedom (d.f.) can be worked out as under:

The degrees of freedom for total sum of squares (TSS) $= (N - 1)$

The degrees of freedom for variance between rows (SSR) $= (r - 1)$

The degrees of freedom for variance between columns(SSC) $= (c - 1)$

The degrees of freedom for residual variance (SSE) $= (c - 1)(r - 1)$

where $c =$ number of columns

and $r =$ number of rows

The degrees of freedom for total sum of squares (TSS) is $(N - 1)$.

Step 10. Mean sum of squares : The mean sum of squares for rows is

$$S_R^2 = \frac{\text{SSR}}{\text{r}-1}$$

The mean sum of squares for columns is

$$S_C^2 = \frac{\text{SSC}}{\text{c}-1}$$

and the mean sum of squares for error is

$$S_E^2 = \frac{\text{SSE}}{(c-1)(r-1)}$$

Step 11. ANOVA Table: The above sum of squares together with their respective degrees of freedom and mean sum of squares can be summarized in the Two-way ANOVA table in the following way.

Source of variation	Sum of squares (SS)	Degrees of freedom (d.f.)	Mean square (MS)	F-ratio
Between rows treatment	SSR	$(r-1)$	$S_R^2 = \frac{SSR}{(r-1)}$	$F_R = \frac{S_R^2}{S_E^2}$
Between columns treatment	SSC	$(c-1)$	$S_C^2 = \frac{SSC}{(c-1)}$	$F_C = \frac{S_C^2}{S_E^2}$
Residual or error	SSE	$(c-1)(r-1)$	$S_E^2 = \frac{SSE}{(c-1)(r-1)}$	
Total	TSS	$(N-1)$		

In the table

c = number of columns

r = number of rows

SSE (residual or error) = TSS(Total)-(SSB +SSC)

.

Step 12. Calculation of F : F-ratio concerning variation between rows is $F_R = \frac{S_R^2}{S_E^2}$

F-ratio concerning variation between rows is $F_C = \frac{S_C^2}{S_E^2}$

In these two cases, if the numerator variance is less than the denominator variance, then numerator and denominator should be interchanged and degrees of freedom should be adjusted accordingly.

Step 13. Conclusion : If F-ratio concerning variation between rows(i.e.F_R) ¿ its table value, then the difference among rows means is considered significant. Similarly, the F-ratio concerning variation between columns can be interpreted.

5.5 Two-way ANOVA technique when repeated values are there:

In case of a two-way design with repeated measurements for all of the categories, we can obtain a separate independent measure of inherent or smallest variations. For this measure we can calculate the sum of squares and degrees of freedom in the same way as we had worked out the sum of squares for variance within samples in the case of one-way ANOVA. Total SS, SS between columns and SS between rows can also be worked out as stated above. We then find left-over sums of squares and left-over degrees of freedom which are used for what is known as 'interaction variation' (Interaction is the measure of inter relationship among the two different classifications). After making all these computations, ANOVA table can be set up for drawing inferences.

Steps involved in Two-way ANOVA technique (when repeated values are there):

The various steps involved are as follows:

Step 1. Define the null hypothesis
H_0 : There is no significant difference between the means on the basis of row factors and there is no significant difference between the means on the basis of column factors.

Step 2. Let n_i be the number of items in ith sample and $N = \sum n_i$=total number of observations.

Step 3. Find the sum of observations in the i th sample,(T_i)
and find the Grand total, $T = \sum T_i = \sum x_{ij}$ = The sum of of all N observations where $i = 1, 2, 3, \ldots r$, where r is the number of rows.

Step 4. Find the correction factor, using

$$\text{CF} = \frac{T^2}{N}$$

Step 5. Find the squares of all observations and then find the total sum of squares

$$\text{TSS} = \sum_i \sum_j (x_{ij})^2 - \text{CF}$$

Step 6. Take the total of different rows, T_i and then obtain the square of each row total (i.e. T_i^2) and divide such squared values of each row by the number

of items in the concerning row(i.e. n_i) and take the total of the result thus obtained. Finally, subtract the correction factor from this total to obtain the sum of squares of deviations for variance between rows (SSR).

$$\text{SSR} = \sum_i \frac{T_i^2}{n_i} - \text{C.F.} \qquad i = 1, 2, 3, \ldots r$$

, where r is the number of rows. (where subscript i represents ith row).

Step 7. Take the total of different columns, T_j and then obtain the square of each column total (i.e. T_j^2) and divide such squared values of each column by the number of items in the concerning column(i.e. n_j) and take the total of the result thus obtained. Finally, subtract the correction factor from this total to obtain the sum of squares of deviations for variance between columns (SSC).

$$\text{SSC} = \sum_j \frac{T_j^2}{n_j} - \text{C.F.} \qquad j = 1, 2, 3, \ldots c,$$

where c is the number of columns.(where subscript j represents jth column).

Step 8. The Sum of squares of deviations for residual or error can be found out by subtracting sum of SSB and SSC from TSS.

SSE= TSS-(SSB+SSC)

i.e. determine the sum of squares within the samples(error variation)

Step 9. Sum of Squares for interrelationship (or interaction) variation can be worked out as under: **SSI=TSS – (SSR + SSC + SSE)**

Step 10. Degrees of freedom (d.f.) can be worked out as under:

The degrees of freedom for total sum of squares (TSS) $= (N - 1)$

The degrees of freedom for variance between rows (SSR) $= (r - 1)$

The degrees of freedom for variance between columns(SSC) $= (c - 1)$

The degrees of freedom for error variance = $\frac{N}{2}$

The degrees of freedom for interaction variation, $SSI = \frac{N}{2} - c - r + 1$

where $c =$ number of columns

and $r =$ number of rows

Step 11. Mean sum of squares : The mean sum of squares for rows is

$$S_R^2 = \frac{\text{SSR}}{\text{r} - 1}$$

The mean sum of squares for columns is

$$S_C^2 = \frac{\text{SSC}}{\text{c} - 1}$$

The mean sum of squares for interaction is

$$S_I^2 = \frac{\text{SSI}}{\frac{N}{2} - c - r + 1}$$

and the mean sum of squares for error is

$$S_E^2 = \frac{\text{SSI}}{\frac{N}{2}}$$

Finally, the ANOVA table for Two Way ANOVA can be set up which as follows:

Step 12. ANOVA Table: The above sum of squares together with their respective degrees of freedom and mean sum of squares can be summarized in the Two-way ANOVA table in the following way.

Source of variation	Sum of Squares(SS)	Degree of freedom	Mean Square (MS)	F-ratio
SS between columns	SSC	$(c-1)$	$S_C^2 = \frac{\text{SSC}}{c-1}$	$F_c = \frac{S_C^2}{S_E^2}$
SS between rows	SSR	$(\text{r}-1)$	$S_R^2 = \frac{SSR}{r-1}$	$F_R = \frac{S_R^2}{S_E^2}$
Interrelationship	SSI	$\frac{N}{2} - c - r + 1$	$S_I^2 = \frac{SSI}{\frac{N}{2}-c-r+1}$	$F_I = \frac{SSI}{S_E^2}$
Within error deviation	SSE	$\frac{N}{2}$	$S_E^2 = \frac{SSE}{\frac{N}{2}}$	
Total	TSS	$N-1$		

In the table

c = number of columns

r = number of rows

Step 13. Calculation of F : F-ratio concerning variation between rows is $F_R = \dfrac{S_R^2}{S_E^2}$

F-ratio concerning variation between rows is $F_C = \dfrac{S_C^2}{S_E^2}$

F-ratio concerning variation between rows is $F_I = \dfrac{S_I^2}{S_E^2}$

In these two cases, if the numerator variance is less than the denominator variance, then numerator and denominator should be interchanged and degrees of freedom should be adjusted accordingly.

Step 14. Conclusion : If F-ratio concerning variation between rows(i.e.F_R) ¿ its table value, then the difference among rows means is considered significant. Similarly, the F-ratio concerning variation between columns can be interpreted.

5.6 CODING METHOD

Coding method is based on an important property of F-ratio that its value does not change if all the n item values are either multiplied or divided by a common figure or if a common figure is either added or subtracted from each of the given n item values. Through this method big figures are reduced in magnitude by division or subtraction and computation work is simplified without any disturbance on the F-ratio. This method should be used specially when given figures are big or otherwise inconvenient. Once the given figures are converted with the help of some common values, then all the steps of the short-cut method (for both ONE-way ANOVA and TWO-way ANOVA) stated above can be adopted for obtaining and interpreting F-ratio.

TWO WAY ANOVA Problems

Problem 137. *The following data represents the number of units of production per day turned out by different workers using 4 different types of machines.*

	Machine Type			
Workers	*A*	*B*	*C*	*D*
1	*44*	*38*	*47*	*36*
2	*46*	*40*	*52*	*43*
3	*34*	*36*	*44*	*32*
4	*43*	*38*	*46*	*33*
5	*38*	*42*	*49*	*39*

1. Test whether the five men differ with respect to mean productivity and
2. Test whether the mean productivity is the same for the four different machine types.

Solution : Let us take the null hypothesis that H_0 : The 5 workers(row factors) do not differ with respect to mean productivity and The mean productivity is the same

for the four different machines(column factors) To simplify calculation let us use coding method and subtract 40 from each value, the new values are

Workers	Machine Type				
	A	B	C	D	Total
1	4	-2	7	-4	5
2	6	0	12	3	21
4	-6	-4	4	-8	-14
4	3	-2	6	-7	0
5	-2	2	9	-1	8
Total	5	-6	38	-17	20

Correction factor , $C.F$

$$= \frac{T^2}{N} = \frac{(20)^2}{20} = 20$$

Total sum of squares, TSS

$$= \sum_i \sum_j (x_{ij})^2 - \text{CF}$$

$$= \left[(4)^2 + (-2)^2 + (7)^2 + (-4)^2 + (6)^2 + (0)^2 + (12)^2 + (3)^2 + (-6)^2 + (-4)^2 + (4)^2 + (-8)^2 + (3)^2 + (-2)^2 + (6)^2 + (-7)^2 + (-2)^2 + (2)^2 + (9)^2 + (-1)^2\right] - 20$$

$$= 574$$

sum of squares of deviations for variance between rows (SSR).

$$\text{SSR} = \sum_i \frac{T_i^2}{n_i} - \text{C.F.}$$

$$= \frac{(5)^2}{4} + \frac{(21)^2}{4} + \frac{(-14)^2}{4} + \frac{(0)^2}{4} + \frac{(8)^2}{4} - 20$$

$$= 181.5 - 20 = 161.5$$

squares of deviations for variance between columns (SSC).

$$\text{SSC} = \sum_j \frac{T_j^2}{n_j} - \text{C.F.}$$

$$= \frac{(5)^2}{5} + \frac{(-6)^2}{5} + \frac{(38)^2}{5} + \frac{(-17)^2}{5} - 20$$

$$= 358.8 - 20 = 338.8$$

The Sum of squares of deviations for residual or error

$$\text{SSE} = TSS - (SSB + SSC)$$

$$= 574 - 161.5 - 338.8 = 73.7$$

Degrees of freedom (d.f.) can be worked out as under:

The degrees of freedom for total sum of squares (TSS) $= (N-1) = 19$

The degrees of freedom for variance between rows (SSR) $= (r-1) = 4$

The degrees of freedom for variance between columns(SSC) $= (c-1) = 3$

The degrees of freedom for residual variance (SSE) $= (c-1)(r-1) = 12$

where $c =$ number of columns

and $r =$ number of rows

Mean sum of squares : The mean sum of squares for rows is

$$S_R^2 = \frac{\text{SSR}}{\text{r}-1} = \frac{161.5}{4} = 40.375$$

The mean sum of squares for columns is

$$S_C^2 = \frac{\text{SSC}}{\text{c}-1} = \frac{338.3}{3} = 112.93$$

and the mean sum of squares for error is

$$S_E^2 = \frac{\text{SSE}}{(c-1)(r-1)} = \frac{73.7}{12} = 6.14$$

ANOVA Table:

Source of variation	Sum of squares (SS)	Degrees of freedom (d.f.)	Mean square (MS)	F-ratio
Between rows treatment	SSR=161.5	4	$S_R^2 = \frac{161.5}{4} = 40.375$	$F_R = \frac{40.375}{6.14} = 6.576$
Between columns treatment	SSC=338.8	3	$S_C^2 = \frac{338.3}{3} = 112.93$	$F_C = \frac{112.93}{6.14} = 18.393$
Residual or error	SSE= 73.7	12	$S_E^2 = \frac{73.7}{12} = 6.14$	
Total	TSS	$(N-1) = 19$		

The table values of F are,

$$F_{0.05}(4, 12) = 3.26 \text{ and } F_{0.05}(3, 12) = 3.49$$

Conclusion: Here, $F_R = 6.576 > F_{0.05}(4, 12) = 3, 26$.

Hence H_0 is rejected for row factor.That is, the 5 workers differ respect to mean productivity. Also, $F_C = 18.393 > F_{0.05}(3, 12) = 3.49$.

Hence H_0 is rejected for column factor.

That is, the mean productivity is not the same for the four machines.

Problem 138. *Set up an analysis of variance table for the following two-way design results:*

Per Acre Production Data of Wheat (in metric tonnes)

Varieties of seeds	***A***	***B***	***C***
Varieties of fertilizers			
W	*6*	*5*	*5*
X	*7*	*5*	*4*
Y	*3*	*3*	*3*
Z	*8*	*7*	*4*

Also state whether variety differences are significant at 5% *level.*

Solution : Refer Class notes

Problem 139. *The following table gives the monthly sales (in thousand rupees) of a certain firm in three states by its four salesmen:*

States	*Salesmen*				*Total*
	A	***B***	***C***	***D***	
X	*5*	*4*	*4*	*7*	*20*
Y	*7*	*8*	*5*	*4*	*24*
Z	*9*	*6*	*6*	*7*	*28*
Total	*21*	*18*	*15*	*18*	*72*

Set up an analysis of variance table for the above information. Calculate $\boldsymbol{F}$*-coefficients and state whether the difference between sales affected by the four salesmen and difference between sales affected in three States are significant.*

Solution : Refer Class notes

Problem 140. *Perform a two-way ANOVA table on the data given below:*

	Treatment 1		
Treatment 2	*1*	*2*	*3*
1	*30*	*26*	*38*
2	*24*	*29*	*28*
3	*33*	*24*	*35*
4	*36*	*31*	*30*
5	*27*	*35*	*33*

Solution : Refer Class notes

Hint:(better to use coding method)

N=15, T=9, CF=5.4, TSS=265.6, SSC=38.8,

SSR=52.9, SSE=173.8, $F_R = 1$=1.642, $F_C = 1.119$

Problem 141. *A company appoints four salesman A, B, C and D observes their sales in three seasons - Summer, winter and Monsoon. The figures (in lac of Rs) are given in the following table:*

Seasons	*Salesman*			
	A	***B***	***C***	***D***
Summer	*45*	*40*	*38*	*37*
Winter	*43*	*41*	*45*	*38*
Monsoon	*39*	*39*	*41*	*41*

Cary out Analysis of variance for two-way classification.

Solution : Refer Class notes

Hint: $F_R = 1.87 < F(6,2) = 19.3$, $F_C = 1 < F(6,3)$, Hence we accept the null hypothesis for both row and column factors. i.e.

That is there is no difference between the sales in the seasons and there is no difference between in the sales of the 4 salesmen.

Problem 142. *Three varieties of coal were analysed by four chemists and the ash-content in the varieties was found to be as under.*

	Chemists			
Varieties	*1*	*2*	*3*	*4*
A	*8*	*5*	*5*	*7*
B	*7*	*6*	*4*	*4*
C	*3*	*6*	*5*	*4*

Carry out the analysis of variance.

Solution : Refer Class notes

Hint: CF=341.33, $\sum x_{ij}^2 = 366$, TSS=24.67, SSR=6.17,

SSC=3.34, SSE=15.16, $S_R^2 = 3.085$, $S_C^2 = 1.113$, $S_E^2 = 2.527$,

$F_R = 1.22$, $F_C = 2.27$,

Two Way ANOVA with Repetitions - Problems:

Problem 143. *Set up ANOVA table for the following information relating to three drugs testing to judge the effectiveness in reducing blood pressure for three different groups of people:*

Amount of Blood Pressure Reduction in Millimeters of Mercury

	Drug		
Group of People	***X***	***Y***	***Z***
A	*14*	*10*	*11*
	15	*9*	*11*
B	*12*	*7*	*10*
	11	*8*	*11*
C	*10*	*11*	*8*
	11	*11*	*7*

Do the drugs act differently?
Are the different groups of people affected differently?
Answer the above questions taking a significant level of 5%.

Solution : Step (i) Here $n_i = 6$ for each row and $n_j = 6$, for each column.
Total no. of observations, $N = 18$
Step (ii) Grand total, $T = \sum T_i = \sum T_j = 187$,
thus, the correction factor, CF $= \frac{T^2}{N} = \frac{187 \times 187}{18} = 1942.72$
Step (iii) Total Sum of squares,

$$\begin{aligned} TSS &= \sum_i \sum_j (x_{ij})^2 - \text{CF} \\ &= \left[(14)^2 + (15)^2 + (12)^2 + (11)^2 + (10)^2 + (11)^2 + (10)^2 \right. \\ &\quad + (9)^2 + (7)^2 + (8)^2 + (11)^2 + (11)^2 + (11)^2 \\ &\quad \left. + (11)^2 + (10)^2 + (11)^2 + (8)^2 + (7)^2\right] - \left[\frac{(187)^2}{18}\right] \\ &= (2019 - 1942.72) \\ &= 76.28 \end{aligned}$$

Step (iv) Sum of squared deviations between rows,

$$SSR = \sum_j \frac{T_i^2}{n_i} - \text{C.F.}$$
$$= \left[\frac{70 \times 70}{6} + \frac{59 \times 59}{6} + \frac{58 \times 58}{6}\right] - 1942.72$$
$$= 816.67 + 580.16 + 560.67 - 1942.72$$
$$= 14.78$$

Step (v) Sum of squared deviations between columns,

$$SSC(i.e., betweendrugs) = \sum_j \frac{T_j^2}{n_j} - \text{C.F.}$$
$$= \left[\frac{73 \times 73}{6} + \frac{56 \times 56}{6} + \frac{58 \times 58}{6}\right] - 1942.72$$
$$= 888.16 + 522.66 + 560.67 - 1942.72$$
$$= 28.77$$

Step (vi) Error Deviations,

$$SSE = \sum (x - \bar{x})^2$$
$$= (14 - 14.5)^2 + (15 - 14.5)^2 + (10 - 9.5)^2 + (9 - 9.5)^2 + (11 - 11)^2$$
$$+ (11 - 11)^2 + (12 - 11.5)^2 + (11 - 11.5)^2 + (7 - 7.5)^2 + (8 - 7.5)^2$$
$$+ (10 - 10.5)^2 + (11 - 10.5)^2 + (10 - 10.5)^2 + (11 - 10.5)^2$$
$$+ (11 - 11)^2 + (11 - 11)^2 + (8 - 7.5)^2 + (7 - 7.5)^2$$
$$= 3.50$$

Step (vii) for interaction variation,

$$SSI = TSS - (SSR + SSC + SSE)$$
$$= 76.28 - [28.77 + 14.78 + 3.50]$$
$$= 29.23$$

Source of variation	SS	d.f.	MS	F-ratio	Table value of F
Between columns (i.e., between drugs)	28.77	$(c-1) = 2$	$\frac{28.77}{2} = 14.385$	$\frac{14.385}{0.389} = 19.0$	$F(2,9) = 4.26$
Between rows (i.e., between people)	14.78	$(r-1) = 2$	$\frac{14.78}{2} = 7.390$	$\frac{7.390}{0.389} = 36.9$	$F(2,9) = 4.26$
Interaction	**29.23**	$\frac{N}{2} - r - c - 1 = 4$	$\frac{29.23}{4} = 7.308$	$\frac{7.308}{0.389} = 18.786$	$F(4,9) = 3.63$
Within (Error)	3.50	$\frac{N}{2} = 9$	$\frac{3.50}{9} = 0.389$		
Total	76.28	$N - 1 = 17$			

The above table shows that all the three F-ratios are significant of 5% level which means that the drugs act differently, different groups of people are affected differently and the interaction term is significant. In fact, if the interaction term happens to be significant, it is pointless to talk about the differences between various treatments i.e., differences between drugs or differences between groups of people in the given case.

Problem 144. *The following table gives the monthly sales (in thousand rupees) of a certain firm in three states by its four salesman. Set up ANOVA table for the information which is given below. Calculate F-coefficients and state whether the difference between sales affected by the four salesman and difference between sales affected in three states are significant. Also taking a significant value of 5%?*

States	*Salesman*			
	A	**B**	**C**	**D**
X	*5*	*4*	*4*	*7*
	3	*5*	*9*	*8*
Y	*7*	*8*	*5*	*4*
	3	*8*	*7*	*5*
Z	*9*	*6*	*6*	*7*
	5	*4*	*3*	*1*

Solution :

States	Salesman				T_i	T_i^2
	A	**B**	**C**	**D**		
X	5	4	4	7	45	2025
	3	5	9	8		
Y	7	8	5	4	47	2209
	3	8	7	5		
Z	9	6	6	7	41	1681
	5	4	3	1		
T_j	9	6	6	1		
T_j^2	81	36	36	1		

Now, the steps involved are as follows:

i. $\mathbf{T = 133, N = 24}$

ii.

$$\text{Correction Factor, CF} = \frac{(\text{ square of } T)}{n} = \frac{(133 \times 133)}{24}$$

$$= 737.04$$

$$= 737$$

iii.

SS total deviation

$$= [25 + 9 + 16 + 25 + 16 + 81 + 49 + 64 + 49 + 9 + 64 + 64 + 25 + 49 + 16 + 25 + 81 + 25 + 36 + 16 + 36 + 9 + 49 + 1] - CF$$

$$= 823 - 737$$

$$= 86$$

iv. SS between columns (i.e. between salesman) deviation

$$= \left[\frac{(32 \times 32)}{6} + \frac{(35 \times 35)}{6} + \frac{(34 \times 34)}{6} + \frac{(32 \times 32)}{6}\right] - \frac{(133 \times 133)}{24}$$

$$= (170.6 + 204.1 + 192.6 + 170.6) - 737.04$$

$$= 737.9 - 737.04$$

$$= 0.86$$

v. SS between rows (i.e. between states) deviation

$$= \left[\frac{(45 \times 45)}{8} + \frac{(47 \times 47)}{8} + \frac{(41 \times 41)}{8}\right] - \frac{(133 \times 133)}{24}$$

$$= (253.1 + 276.1 + 210.1) - 737.04$$

$$= 739.3 - 737.04$$

$$= 2.26$$

(vi) SS within sales (i.e. error) deviation

$$SSE = (1 + 1 + 0.25 + 0.25 + 6.25 + 6.25 + 0.25 + 0.25 + 4 + 4 + 0 + 0 + 1 + 1 + 0.25 + 0.25 + 4 + 4 + 1 + 1 + 2.25 + 2.25 + 9 + 9)$$

$$= 58.5$$

vii) SS for interrelationship variation,

$$= 86 - (0.86 + 2.26 + 58.5) = 24.38$$

Source of variation	Sum of Squares (SS)	Degree of freedom (DOF)	Mean Square (MS)	F-ratio
SS between columns	0.86	$(4-1) = 3$	$\frac{0.86}{3.28} = 0.28$	$\frac{4.87}{0.28=17.39} = 0.057$
SS between rows	2.26	$(3-1) = 2$	$\frac{2.26}{2} = 1.13$	$\frac{4.87}{1.13} = 4.31$
SS of Interrelationship	24.38	4	$\frac{24.38}{4} = 6.0$	$\frac{6.0}{4.87} = 1.23$
SS within sales(error)	58.5	$(24-12) = 12$	$\frac{58.5}{12} = 4.87$	
Total	86	$(24-1) = 23$		

(Compare with table value and write your conclusion!)

5.7 ANOVA IN LATIN-SQUARE DESIGN

Latin-square design is an experimental design used frequently in agricultural research. In such a design the treatments are so allocated among the plots that no treatment occurs, more than once in any one row or any one column. The ANOVA technique in case of Latin-square design remains more or less the same as we have already stated in case of a two-way design, excepting the fact that the variance can be split into four parts as under:
(i) variance between columns;
(ii) variance between rows;
(iii) variance between varieties;
(iv) residual variance.
All these above stated variances are worked out as under:

Step 1. Define the null hypothesis

Step 2. find $N = \sum n_i = \sum n_j$=total number of observations.

Step 3. Find the sum of observations in the i th row,(T_i) and find the Grand total, $T = \sum T_i = \sum x_{ij}$

Step 4. Find the correction factor, using

$$\text{CF} = \frac{T^2}{N}$$

Step 5. Find the squares of all observations and then find the total sum of squares,

$$\text{TSS} = \sum_i \sum_j (x_{ij})^2 - \text{CF}$$

Step 6. Find the sum of squares of deviations for variance between rows (SSR).

$$\text{SSR} = \sum_i \frac{T_i^2}{n_i} - \text{C.F.}$$

where T_i is the sum of observations in the i th row,

Step 7. Find the sum of squares of deviations for variance between columns (SSC).

$$SS = sum\frac{(T_j)^2}{n_j} - CF$$

where T_j is the sum of observations in the j th column,

Step 8. Find the sum of squares of deviations for variance between varieties(letters) (SSV or SSL).

$$SSV \text{ or } SSL = \sum \frac{(T_v)^2}{n_v} - CF$$

where T_v is the sum of observations of variety type v or letter types v,

Step 9. Find the Sum of squares for residual variance(error),

$$SSE = TSS - (SSR + SSC + SSV)$$

Step 10. Degrees of freedom (d.f.) can be worked out as under:

The degrees of freedom for total sum of squares (TSS) $= (N-1)$

The degrees of freedom for variance between rows (SSR) $= (r-1)$

The degrees of freedom for variance between columns(SSC) $= (c-1)$

The degrees of freedom for variance between varieties/Letters(SSL) $= (v-1)$

The degrees of freedom for residual variance (SSE) $= (c-1)(c-2)$

where c = number of columns, r = number of rows and v = number of varieties In place of c we can as well write r or v since in Latin-square design $c = r = v$.

ANOVA table can now be set up as shown below:

Source of variation	SS	d.f.	MS	F-ratio
Between columns	SSC	c-1	$S_C^2 = \frac{SSC}{c-1}$	$F_C = \frac{S_C^2}{S_E^2}$
Between rows	SSR	r-1	$S_R^2 = \frac{SSR}{r-1}$	$F_R = \frac{S_R^2}{S_E^2}$
Between varieties	SSL	c-1	$S_L^2 = \frac{SSL}{c-1}$	$F_L = \frac{S_L^2}{S_E^2}$
Residual or error	SSE	(c-1)(C-2)	$S_E^2 = \frac{SSE}{(c-1)(c-2)}$	
Total	TSS	N-1		

Step 11. Conclusion can be drawn based on the calculated values of F compared with tabulated value of F.

Problem 145. *Analyse and interpret the following statistics concerning output of wheat per field obtained as a result of experiment conducted to test four varieties of wheat viz., A, B, C and D under a Latinsquare design.*

C 25	***B*** 23	*A* 20	***D*** 20
A 19	***D*** 19	*C* 21	***B*** 18
B 19	***A*** 14	***D*** 17	*c* 20
D 17	*c* 20	***B*** 21	***A*** 15

Solution: Using the coding method, let us subtract 20 from the figures given in each of the small squares and obtain the coded figures as under:

C 5	***B*** 3	A 0	***D*** 0
A -1	***D*** -1	C 1	***B*** -2
B -1	***A*** -6	***D*** -3	c 0
D -3	C -3	***B*** 1	***A*** -5

row \ column 1	2	3	4	T_i	T_i^2	
1	***C*** 5	***B*** 3	A 0	***D*** 0	8	64
2	***A*** -1	***D*** -1	C 1	***B*** -2	-2	4
3	***B*** -1	***A*** -6	***D*** -3	C 0	-10	100
4	***D*** -3	C -3	***B*** 1	***A*** -5	-7	49
T_j	0	-4	-1	-7	T=-12	
T_j^2	0	16	1	49		

Squaring the coded values in various columns and rows we have the following table of square terms:

C 25	B 9	A 0	D 0
A 1	D 1	C 1	B 4
B 1	A 36	D 9	c 0
D	C	B	A
9	0	1	25

$$\text{Correction factor , } CF = \frac{(T)^2}{n} = \frac{(-12)(-12)}{16} = 9$$

$$SS \text{ for total variance } = \Sigma (X_{ij})^2 - \frac{(T)^2}{n} = 122 - 9 = 113$$

$$SS \text{ for variance between columns } = \sum \frac{(T_j)^2}{n_j} - CF$$

$$= \left\{ \frac{(0)^2}{4} + \frac{(-4)^2}{4} + \frac{(-1)^2}{4} + \frac{(-7)^2}{4} \right\} - 9$$

$$= \frac{66}{4} - 9 = 7.5$$

SS for variance between rows

$$= \sum \frac{(T_i)^2}{n_i} - CF$$

$$= \left\{ \frac{(8)^2}{4} + \frac{(-3)^2}{4} + \frac{(-10)^2}{4} + \frac{(-7)^2}{4} \right\} - 9$$

$$= \frac{222}{4} - 9 = 46.5$$

SS for variance between varieties would be worked out as under:

For finding SS for variance between varieties, we would first rearrange the coded data in the following form:

Varieties of wheat	Yeld in different parts of field				Total (T_V)
	I	II	III	IV	
A	-1	-6	0	-5	-12
B	-1	3	1	-2	1
C	5	0	1	0	6
D	-3	-1	-3	0	-7

Now we can work out SS for variance between varieties as under:

$$\text{SS for variance between varieties } = \sum \frac{(T_v)^2}{n_v} - CF$$

$$= \left\{ \frac{(-12)^2}{4} + \frac{(1)^2}{4} + \frac{(6)^2}{4} + \frac{(-7)^2}{4} \right\} - 9$$

$$= \frac{230}{4} - 9 = 48.5$$

$\therefore$ Sum of squares for residual variance will work out to

$$113 - (7.5 + 46.5 + 48.5) = 10.50$$

d.f. for variance between columns $= (c-1) = (4-1) = 3$ d.f. for variance between rows $= (r-1) = (4-1) = 3$ d.f. for variance between varieties $= (v-1) = (4-1) = 3$ d.f. for total variance

$$= (n-1) = (16-1) = 15$$

d.f. for residual variance

$$= (c-1)(c-2) = (4-1)(4-2) = 6$$

ANOVA table can now be set up as shown below:

Source of variation	SS	d.f.	MS	F-ratio	5% F-limit
Between columns	7.50	3	$\frac{7.50}{3} = 2.50$	$\frac{2.50}{1.75} = 1.43$	$F(3,6) = 4.76$
Between rows	46.50	3	$\frac{46.50}{3} = 15.50$	$\frac{15.50}{1.75} = 8.85$	$F(3,6) = 4.76$
Between varieties	48.50	3	$\frac{48.50}{3} = 16.17$	$\frac{16.17}{1.75} = 9.24$	$F(3,6) = 4.76$
Residual or error	10.50	6	$\frac{10.50}{6} = 1.75$		
Total	113.00	15			

The above table shows that variance between rows and variance between varieties are significant and not due to chance factor at 5% level of significance as the calculated values of the said two variances are 8.85 and 9.24 respectively which are greater than the table value of 4.76 . But variance between columns is insignificant and is due to chance because the calculated value of 1.43 is less than the table value of 4.76 .

Problem 146. *Analyse the variance in the following latin square of yields (in kos) of paddy where A, B, C, D denote the different methods of Cultivation.*

D_{122}	A_{121}	C_{123}	B_{122}
B_{124}	C_{123}	A_{122}	D_{125}
A_{120}	B_{119}	D_{120}	C_{121}
C_{122}	D_{123}	B_{121}	A_{122}

Examine whether the different methods of cultivation have given significantly different yields. Given that $F_{3,6} = 4.76$.

Solution : i) There is no difference between sows.
ii) There is no difference between colums.
iii) There is no difference between letters (methods of cultivation) Using coding Method, subtract by 120 $N = 4 + 4 + 4 + 4 = 16$

D_{2}	A_{1}	C_{3}	B_{2}
B_{4}	C_{3}	A_{2}	D_{5}
A_{0}	B_{-1}	D_{0}	C_{1}
C_{2}	D_{3}	B_{1}	A_{2}

Correction factor $= \frac{T^2}{\eta} = \frac{(30)^2}{16} = 56.25$
TSS $= \sum x^2 - \frac{T^2}{n} = 92 - 56.25 = 35.75$
$SSR = \frac{(8)^2 + (14)^2 + (0)^2 + (8)^2}{} - 56.25 = 24.75$
$SSC = \frac{(8)^2+(6)^2+(6)^2+(10)^2}{4} - 56.25 = 2.75$
From letters [Take values with respective letters]

$$\sum A = 1 + 2 + D + 2 = 5, \quad \sum C = 3 + 3 + 1 + 2 = 9$$
$$\sum B = 2 + 4 + (-1) + 1 = 6, \quad \sum D = 2 + 5 + 0 + 3 = 10.$$

SSL(varieties)

$$= \frac{(5)^2 + (6)^2 + (9)^2 + (10)^2}{4} - 56.25 = 4.25.$$

$$\text{SSE} = TSS - SSR - SSC - SSV$$
$$\text{SSE} = 35.75 - 24.75 - 2.75 - 4.25 = 4.$$

Source of variation	SS	d.f.	MS	F-ratio	5% F-limit
Between columns	2.75	3	0.92		1.37
Between rows	24.75	3	8.25	12.31	$F(3,6) = 4.76$
Between varieties	4.25	3	1.42	2.12	$F(3,6) = 4.76$
Residual or error	4	6	0.67		
Total	35.75	15			

$\therefore$ Hypothesis accepted for columns & letters (cultivation method), and Hypothesis rejected for rows.

Question Bank

1. A test was given to five students taken at random from the fifth class of three schools of a town. The individual scores are

School I	9	7	6	5	8
School II	7	4	5	4	5
School III	6	5	6	7	6

   Carry out the analysis of variance.

2. The following figures relate to production in kg of three varieties A, B and C of wheat shown in 12 plots.

   A: 20 18 19
   B: 17 16 19 18
   C: 20 21 20 19 18

   Is there any significant difference in the production of the three varieties Ans : Calculated $F = 9.11 <$ Table value of $F(9,2) = 19.3$

3. Determine if there is a difference in the mean daily calcium intake for people with normal bone density, osteopenia, and osteoporosis at a 0.05 alpha level. The data was recorded as follows:

Normal Density	Osteopenia	Osteoporosis
1200	1000	890
1000	1100	650
980	700	1100
900	800	900
750	500	400
800	700	350

Hint: $F = 1.395 < F(0.05, 2, 15) = 3.68$, accept

4. Three processes A, B and C are tested to see whether their outputs are equivalent. The following observations of outputs are made:

A	10	12	13	11	10	14	15	13
B	9	11	10	12	13			
C	11	10	15	14	12	13		

Carry out the analysis of variance and state your conclusion.

5. Suppose in an industrial experiment that an engineer is interested in how the mean absorption of moisture in concrete varies among 5 different concrete aggregates. The samples are exposed to moisture for 48 hours. It is decided that 6 samples are to be tested for each aggregate, requiring a total of 30 samples to be tested. The data are recorded in the following Table:

Aggregate:	**1**	**2**	**3**	**4**	**5**	
	551	**595**	**639**	**417**	**563**	
	457	**580**	**615**	**449**	**631**	
	450	**508**	**511**	**517**	**522**	
	731	**583**	**573**	**438**	**613**	
	499	**633**	**648**	**415**	**656**	
	632	**517**	**677**	**555**	**679**	
Total	**3320**	**3416**	**3663**	**2791**	**3664**	**16, 854**

6. As head of the department of a consumers research organization you have the responsibility of testing and comparing life times of 4 brands of electric bulbs.suppose you test the life time of 3 electric bulbs each of 4 brands,the data is given below,each entry representing the life time of an electric bulb,measured

in hundreds of hours.

A	B	C	D
20	**25**	**24**	**23**
19	**23**	**20**	**20**
21	**21**	**22**	**20**

7. Set up an analysis of variance table for the following per acre production data for three varieties of wheat, each grown on 4 plots and state if the variety differences are significant.

Plot of land	Per acre production data		
	A Variety of wheat		
	A	***B***	***C***
1	6	5	5
2	7	5	4
3	3	3	3
4	8	7	4

Hint : $F = 1.5 < F(2, 9) = 4.26$

8. The varieties of wheat A, B, C were sown in 4 plots each and the following yields in quintals per acre were obtained.

A	8	4	6	7
B	7	6	5	3
C	2	5	4	4

Test the significance of difference between the yields of varieties, given that 5% tabulated value of F for 2 and 9 degrees of freedom is 4.26 .

9. Three types of fertilizers are used on three groups of plants for 5 weeks. We want to check if there is a difference in the mean growth of each group. Using the data given below apply a one-way ANOVA test at 0.05 significant level

Fertilizer-1	6	8	4	5	3	4
Fertilizer-2	8	12	9	11	6	8
Fertilizer-3	13	9	11	8	7	12

Hint: CF=1152, TSS=152, SSR=84, SSE=68, $F = 9.33 > F(0.05, 2, 15) = 3.68$, reject

10. Three different machines are used for a production. On the basis of the outputs, set up one - way ANOVA table and test whether the machines are equally effective.

Outputs		
Machine I	Machine II	Machine III
10	9	20
15	7	16
11	5	10
10	6	14

Given that the value of F at 5% level of significance for $(2, 9)$d. f is 4.26

11. A manufacturing company has purchased three new machines of different makes and wishes to determine whether one of them is faster than the others in Producing a certain output. Five hourly production figures are observed at random from each other machine and the results are given below:

Observation	A_1	A_2	A_3
1	25	31	24
2	30	39	30
3	36	38	28
4	38	42	25
5	31	35	28

Use ANOVA and determine whether the machines are significantly different in their mean speed.

12. Three different kinds of food are tested on three groups of rats for 5 weeks. The objective is to check the difference in mean weight (in grams) of the rats per week. Apply one-way ANOVA using a 0.05 significance level to the following data:

Food 1	8	12	19	8	6	11
Food 2	4	5	4	6	9	7
Food 3	11	8	7	13	7	9

Hint: CF=1317.55, TSS=228.45, SSB=84, SSE=68, $S_1^2 = 42, S_2^2 = 4.53, F = 9.33$, Reject

13. A trial was run to check the effects of different diets. Positive numbers indicate weight loss and negative numbers indicate weight gain. Check if there is an average difference in the weight of people following different diets using an ANOVA Table.

Low Fat	Low Calorie	Low protein	Low carbohydrate
8	2	3	2
9	4	5	2
6	3	4	-1
7	5	2	0
3	1	3	3

Hint : CF=252,, TSS=123.2, SSR=75.8, SSE=47.4, $F = 8.43 > F(0.05, 3, 16) = 3.24$ reject

14. The following data represents the number of units of production per day turned out by different workers using 4 different types of machines.
The following data represents the number of units of production per day turned out by different workers using 4 different types of machines.

Workers	Machine Type A	B	C	D
1	44	38	47	36
2	46	40	52	43
3	34	36	44	32
4	43	38	46	33
5	38	42	49	39

1. Test whether the five men differ with respect to mean productivity and
2. Test whether the mean productivity is the same for the four different machine types.

Hint: Using coding, Subtract 40 from each value, CF=20, TSS=574,SSC=161.5, SSR=338.3, SSE=73.7,

$F_R = 6.576. F_c = 18.393$, H_0 is rejected for both row and column factors

15. Set up an analysis of variance table for the following two-way design results: Per Acre Production Data of Wheat (in metric tonnes)

Varieties of seeds	A	B	C
Varieties of fertilizers			
W	6	5	5
X	7	5	4
Y	3	3	3
Z	8	7	4

Also state whether variety differences are significant at 5% level.

16. The following data show the number of worms quarantined from the GI areas of four groups of muskrats in a carbon tetrachloride anthelmintic study. Conduct a two-way ANOVA test.

I	II	III	IV
33	41	12	38
32	38	35	43
26	40	46	25
14	23	22	13
30	21	11	26

Hint: CF=48, TSS=2245, SSC=151, SSR=1056.25, SSE=1037.75, $F_R = 3.053, F_c = 1.718$, Accept

17. The following table gives the monthly sales (in thousand rupees) of a certain firm in three states by its four salesmen:

States	Salesmen				Total
	A	B	C	D	
X	5	4	4	7	20
Y	7	8	5	4	24
Z	9	6	6	7	28
Total	21	18	15	18	72

Set up an analysis of variance table for the above information. Calculate F-coefficients and state whether the difference between sales affected by the four salesmen and difference between sales affected in three States are significant.

18. A company appoints four salesman A, B, C and D observes their sales in three seasons - Summer, winter and Monsoon. The figures (in lac of Rs) are given in the following table:

Seasons	Salesman			
	A	B	C	D
Summer	45	40	38	37
Winter	43	41	45	38
Monsoon	39	39	41	41

Cary out Analysis of variance for two-way classification.

19. Three varieties of coal were analysed by four chemists and the ash-content in the varieties was found to be as under.

	Chemists			
Varieties	1	2	3	4
A	8	5	5	7
B	7	6	4	4
C	3	6	5	4

Carry out the analysis of variance.

Hint: CF=341.33, TSS=24.67, SSR=6.17, SSC=3.34, SSE=15.16, $F_R = 1.22, F_c = 3.27$, Accept H_0 for both row and column factors.

20. Set up ANOVA table for the following information relating to three drugs testing to judge the effectiveness in reducing blood pressure for three different groups of people:

Amount of Blood Pressure Reduction in Millimeters of Mercury

	Drug		
Group of People	X	Y	Z
A	14	10	11
	15	9	11
B	12	7	10
	11	8	11
C	10	11	8
	11	11	7

Do the drugs act differently? Are the different groups of people affected differently?Answer the above questions taking a significant level of 5%.

21. The following table gives the monthly sales (in thousand rupees) of a certain firm in three states by its four salesman. Set up ANOVA table for the information which is given below. Calculate F-coefficients and state whether the

difference between sales affected by the four salesman and difference between sales affected in three states are significant. Also taking a significant value of 5%?

States	Salesman			
	A	B	C	D
X	5	4	4	7
	3	5	9	8
Y	7	8	5	4
	3	8	7	5
Z	9	6	6	7
	5	4	3	1

22. The following are the number of mistakes made in 5 successive days by 4 technicians working for a photographic laboratory. Test whether the difference among the four sample means can be attributed to chance. [Test at a level of significance $\alpha = 0.01$].

I	II	III	IV
6	14	10	9
14	9	12	12
10	12	7	8
8	10	15	10
11	14	11	11

Hints: T=213, CF=2268.45, TSS=114.55, SSR=12.95, SSE=101.6, $F = 1.47 < F(16, 3) = 8.7$, Accept

23. The following data represents the number of units of production per day turned out by different workers using 4 different types of machines.

Workers	Machine Type			
	A	B	C	D
1	44	38	47	36
2	46	40	52	43
3	34	36	44	32
4	43	38	46	33
5	38	42	49	39

1. Test whether the five men differ with respect to mean productivity and
2. Test whether the mean productivity is the same for the four different machine types.

Hint: $F_R = 6.576 > F_{0.05}(4, 12) = 3,26$, $F_C = 18.393 > F_{0.05}(3, 12) = 3.49$.

Hence H_0 is rejected for both row and factors.

24. Set up the analysis of variance for the following results of a Latin Square Design. Use 0.01 level of significance.

A 12	C 19	B 10	D 8
C 18	B 12	D 6	A 7
B 22	D 10	A 5	C 21
D 12	A 7	C 27	B 17

Ans: $F_R = 1.12, F_c = 1.08, F_v = 11.72$.

The values F_R and F_C are less than the table value $F_{0.01}(3, 6) = 9.78$ and the value F_V is greater than the table value $F_{0.01}(3, 6) = 9.78$

so we conclude that there is no significant difference due to rows and columns but there is significant difference due to the treatments(letters).

25. The figures in the following $5 * 5$ Latin square are the numbers of minutes, engines E_1, E_2, E_3, E_4 and E_5 tuned up by mechanics M_1, M_2, M_3, M_4 and M_5, ran with a gallon of fuel A, B, C, D and E.

	E_1	E_2	E_3	E_4	E_5
M_1	A	B	C	D	E
	31	24	20	20	18
M_2	B	C	D	E	A
	21	27	23	25	31
M_3	C	D	E	A	B
	21	27	25	29	21
M_4	D	E	A	B	C
	21	25	33	25	22
M_5	E	A	B	C	D
	21	37	24	24	20

Use the level of significance $\alpha = 0.01$ to test 1. The null hypothesis H_0 that there is no difference in the performance of the five engines.

2. H_0 that the persons who tuned up these engines have no effect on their performance.

3. H_0 that the engines perform equally well with each of the fuels.

(Ans: $F_R = 2.31, F_c = 8.24, F_L = 31.28$.

The values F_R and F_C are less than the table value $F_{0.01}(4, 12) = 5.41$ and the value F_V is greater than the table value $F_{0.01}(4, 12) = 5.41$

So we conclude that there is no significant difference due to rows (Mechanics) and columns(Engines), but there is significant difference due to the fuels(letters)).

26. In a Latin square experiment, given below are the yields in quintals per acre on paddy crop carried out for testing the effect of five fertilizers A, B, C, D, E. Analyze the data for variations.

B 25	A 18	E 27	D 30	C 27
A 19	D 31	C 29	E 26	B 23
C 28	B 22	D 33	A 18	E 27
E 28	C 26	A 20	B 25	D 33
D 32	E 25	B 23	C 28	A 20

Hint: Use coding, subtract 25 from each value

N=28, T=18, CF=12.96, TSS=483.04, SSR=3.04, SSC=14.24, SSL=454.64, SSE=11.12, $F_R = 1.22, F_C = 3.83, F_L = 122.22$, Write the conclusion

27. Present your conclusions after doing analysis of variance to the following results of the Latin-square design experiment conducted in respect of five fertilizers which were used on plots of different fertility.

A 16	**B** 10	**C** 11	**D** 9	**E** 9
E 10	**C** 9	**A** 14	**B** 12	**D** 11
B 15	**D** 8	**E** 8	**C** 10	**A** 18
D 12	**E** 6	**B** 13	**A** 13	**C** 12
C 13	**A** 11	**D** 10	**E** 7	**B** 14

Hint: Subtracting 10 from each value (coding),

CF=38.44, TSS=196.56, SSR=2.16, SSC=66.56,SSL=122.56, SSE=5.28, $F_R = 1.227, F_c = 37.81, F_L = 69.63$, Compare, H_0 is accepted for row factors, but rejected for Column factors and Letters

28. Analyze and interpret the following statistics concerning output of wheat per field obtained as a result of experiment conducted to test four varieties of wheat viz. A, B, C and D under a Latin- square design

C 25	B 23	A 20	D 20
A 19	D 19	C 21	B 18
B 19	A 14	D 17	C 20
D 17	C 20	B 21	A 15

Hint: Subtracting 20 from each value (coding),

CF=9, TSS=113, SSR=4, SSC=7.5, SSL=48.5, SSE=10.5, $F_R = 8.85, F_c = 1.428, F_L = 9.23$, Compare, H_0 is accepted for column factors, but rejected for row factors and Letters