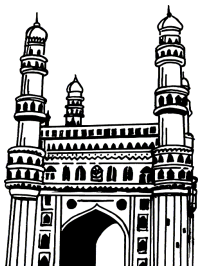
Rahul's ✔
Topper's Voice



M.C.A.

I Year I Sem

(Osmania University)

Latest 2024 Edition

PROBABILITY & STATISTICS

- ☞ **Study Manual**
- ☞ **FAQ's and Important Questions**
- ☞ **Solved Model Papers**
- ☞ **Solved Previous Question Papers**

- by -

WELL EXPERIENCED LECTURER

Rahul Publications™
Hyderabad. Cell : 9391018098, 9505799122

M.C.A.

I Year I Sem

(Osmania University)

PROBABILITY & STATISTICS

Inspite of many efforts taken to present this book without errors, some errors might have crept in. Therefore we do not take any legal responsibility for such errors and omissions. However, if they are brought to our notice, they will be corrected in the next edition.

Price `. 209 -00

Sole Distributors : **Cell : 9391018098, 9505799122**

VASU BOOK CENTRE

Shop No. 2, Beside Gokul Chat, Koti, Hyderabad.

Maternity Hospital Opp. Lane, Narayan Naik Complex, Koti, Hyderabad.

Near Andhra Bank, Subway, Sultan Bazar, Koti, Hyderabad -195.

PROBABILITY & STATISTICS

CONTENTS

STUDY MANUAL

FAQ's and Important Questions	**III - VIII**
Unit - I	**1 - 26**
Unit - II	**27 - 64**
Unit - III	**65 - 89**
Unit - IV	**90 - 116**
Unit - V	**117 - 164**

SOLVED MODEL PAPERS

Model Paper - I	**165 - 166**
Model Paper - II	**167 - 168**
Model Paper - III	**169 - 170**

SOLVED PREVIOUS QUESTION PAPERS

August - 2021	**171 - 177**
April / May - 2023	**178 - 190**
October / November - 2023	**191 - 202**

SYLLABUS

UNIT - I

Vector Spaces - Vector Spaces and Subspaces - Null Spaces, Column Spaces and Linear Transformations. Linearly Independent Sets - Bases - Coordinate Systems.

UNIT - II

Probability - Basic terminology, Three types of probability, Probability rules, Statistical independence, statistical dependency, Bayes' theorem.

Probability Distributions - Random variables, expected values, binomial distribution, Poisson distribution, normal distribution, choosing correct distribution.

UNIT - III

Sampling and Sampling Distributions - Random sampling, Non-Random Sampling distributions, operational considerations in sampling.

Estimation - Point estimates, interval estimates, confidence intervals, calculating interval estimates of the mean and proportion, t-distribution, determination of sample size in estimation.

UNIT - IV

Testing Hypothesis - one sample tests - Hypothesis testing of mean when the population standard deviation is known, powers of hypothesis test, hypotheses testing of proportions, hypotheses testing of means when standard deviation is not known.

Testing Hypotheses - Two sample tests - Tests for difference between means - large sample, small sample, with dependent samples, testing for difference between proportions – Large sample.

UNIT - V

Chi-square and Analysis of Variance - Chi-square as test of independence, chi-square as a test of goodness of fit, analysis of variance, inferences about a population variance, inferences about two population variances.

Regression and Correlation - Simple Regression - Estimation using regression line, correlation analysis, making inferences about population parameters, limitations, errors and caveats in regression and correlation analysis. Multiple Regression and correlation analysis. Finding multiple regression equations and making inferences about population parameters.

Contents

UNIT - I

Topic		Page No.
1.1	Vector Spaces and Subspaces	1
1.2	Null Spaces, Column Spaces and Linear Transformations	5
1.3	Linearly Independent Sets, Bases	11
1.4	Coordinate Systems	16

UNIT - II

2.1	Probability	27
	2.1.1 Basic terminology	27
	2.1.2 Three Types of Probability	29
	2.1.3 Probability Rules	32
2.2	Probabilities under Conditions of Statistical Independence	36
2.3	Probabilities under Conditions of Statistical Dependency	40
2.4	Bayes' Theorem	42
2.5	Random Variables, Expected Values	45
2.6	Binomial Distribution	51
2.7	Poisson Distribution	55
2.8	Normal Distribution	58
2.9	Choosing Correct Probability Distribution	64

UNIT - III

3.1	Sampling	65
	3.1.1 Random Sampling, Non-Random Sampling Distributions	67
3.2	Operational Considerations in Sampling	68
3.3	Sampling Distributions	69
3.4	Estimation	69
	3.4.1 Types	70
	3.4.1.1 Point estimates, interval estimates	70
	3.4.2 Confidence Intervals	72
	3.4.3 Calculating Interval Estimates of the mean and Proportion	73
	3.4.4 Determination of Sample Size in Estimation	74
3.5	T-Distribution	78

Topic **Page No.**

UNIT - IV

4.1 Testing Hypothesis 90

4.1.1 One Sample Tests 92

4.1.2 Hypothesis Testing of Mean when the Population Standard Deviation is known 93

4.1.3 Powers of Hypothesis Test 95

4.1.4 Hypothesis Testing of Proportions 95

4.1.5 Hypothesis Testing of means when Standard Deviation is not known. 101

4.2 Two Sample Tests 104

4.3 Tests for Difference between means - Large Sample 107

4.3.1 Dependent samples 109

4.3.2 Testing for difference between Proportions – Large Sample 109

UNIT - V

5.1 Chi-square 117

5.1.1 Chi-square as Test of Independence 118

5.1.2 Chi-square as a Test of Goodness of Fit 118

5.2 Analysis of Variance 124

5.2.1 Inferences about a Population Variance 124

5.2.2 Inferences about Two Population Variances 139

5.3 Correlation 143

5.4 Regression 150

5.4.1 Limitations 151

5.4.2 Estimation using Regression Line 152

5.4.3 Making inferences about population parameters 156

5.5 Errors and Caveats in Regression and Correlation Analysis 157

5.6 Multiple Regression 159

5.6.1 Finding multiple regression equations 159

5.6.2 Making Inferences about Population Parameters 162

Frequently Asked and Important Questions

UNIT - I

1. Explain about Vector Spaces and Subspaces. **(May-23, Aug.-21, Imp.)**

Ans :

Refer to Unit-I, Page No. 1, Q.No. 1

---

2. $A = \begin{bmatrix} 1 & -3 & -2 \\ -5 & 9 & 1 \end{bmatrix}$ $U = \begin{bmatrix} 5 \\ 3 \\ -2 \end{bmatrix}$ **Determine if U belongs to the Null space of A.**

Sol : **(Imp.)**

Refer to Unit-I, Page No. 5, Prob. 7

---

3. Find A such that the set $\left\{ \begin{bmatrix} b-c \\ 2b+3d \\ b+3c-3d \\ c+d \end{bmatrix} \right\}$ **is Col A**

Sol : **(Imp.)**

Refer to Unit-I, Page No. 8, Prob. 15

---

4. Find a basis for the set of vectors in R^3 in the plane $x + 2y + z = 0$.

Sol : **(Imp.)**

Refer to Unit-I, Page No. 14, Prob. 22

---

5. $H = \text{Span}\{V_1, V_2, V_3\}$ **&** $B = (V_1\ V_2\ V_3)$ **Show that B is a basis for H & X is in H, find the B-coordinate vector of X fee.**

$$V_1 = \begin{bmatrix} -6 \\ 4 \\ -9 \\ 4 \end{bmatrix} V_2 = \begin{bmatrix} 8 \\ -3 \\ 7 \\ -3 \end{bmatrix} V_3 = \begin{bmatrix} -9 \\ 5 \\ -8 \\ 3 \end{bmatrix} X = \begin{bmatrix} 4 \\ 7 \\ -8 \\ 3 \end{bmatrix}$$

Sol : **(Imp.)**

Refer to Unit-I, Page No. 20, Prob. 31

6. Find the coordinates of (6i, 7, 8i) w.r.t the basis s = {(1, 0, 0) (1, 1, 0) (1, 1, 1)} of $c^3(c)$.

Sol : **(Imp.)**

Refer to Unit-I, Page No. 24, Prob.35

UNIT - II

1. Explain the basic terminology are used in probability.

Ans : **(Imp.)**

Refer to Unit-II, Page No. 27, Q.No. 2

2. Explain addition theorem of probability.

Ans : **(Imp.)**

Refer to Unit-II, Page No. 32, Q.No. 4

3. Write about the various probabilities under statistical independence.

Ans : **(Imp.)**

Refer to Unit-II, Page No. 36, Q.No. 5

4. State the explain Bayes' theorem.

Ans : **(May-23, Imp.)**

Refer to Unit-II, Page No. 42, Q.No. 8

5. First box contains 2 black, 3 red, 1 white balls, second box contains 1 black, 1 red, 2 white balls and third box contains 5 black, 3 red, 4 white balls. Of these a box is selected at random. From it a red ball is randomly drawn. If the ball is red, find the probability that it is from second box.

Sol : **(Imp.)**

Refer to Unit-II, Page No. 44, Prob. 12

6. In a bolt factory machines A, B, C manufacture 20%, 30% and 50% of the total of their output and 6%, 3% and 2% are defective. A bolt is drawn at random and found to be defective. Find the probabilities that it is manufactured from

(i) Machine A

(ii) Machine B

(iii) Machine C

Sol : **(Imp.)**

Refer to Unit-II, Page No. 44, Prob. 13

7. What is a binomial distribution? Explain its properties and applications.

Ans : **(Imp.)**

Refer to Unit-II, Page No. 51, Q.No. 11

8. Explain briefly about Poisson Distribution and Applications.

Ans : **(Imp.)**

Refer to Unit-II, Page No. 55, Q.No. 12

9. A large flashlight is powered by 5 batteries. Suppose that the life of a battery is normally distributed with m = 150 hours and s = 15 hours. The flashlight will cease functioning if one or more of its batteries go dead. Assuming the lives of batteries are independent, what is the probability that flashlight will operate more than 130 hours ?

Sol : **(Imp.)**

Refer to Unit-II, Page No. 62, Prob.24

UNIT - III

1. State the Basic terminology used in sampling.

Ans : **(Imp.)**

Refer to Unit-III, Page No. 65, Q.No. 2.

2. Explain different types of sampling methods.

Ans : **(May-23, Imp.)**

Refer to Unit-III, Page No. 67, Q.No. 5.

3. What is Sampling Distributions? Explain the characteristics of Sampling Distributions.

Ans : **(Imp.)**

Refer to Unit-III, Page No. 69, Q.No. 7

4. Explain briefly about various types of estimates.

Ans : **(Nov.-23, May-23, Imp.)**

Refer to Unit-III, Page No. 70, Q.No. 10

5. What are the properties of a good estimator?

Ans : **(Imp.)**

Refer to Unit-III, Page No. 71, Q.No. 11

6. A manufacturer of certain electric Bulbs claims that is bulbs have mean life 25 months with standard deviation 5 months a random sample of 6 such bulbs areas follows.

Life of Bulbs	1	2	3	4	5	6
Months	24	26	30	20	20	18

Can you regard the procedure clamied to be valid at 1% Los ($t_{0.01}$ = 4.032).

Sol : **(Imp.)**

Refer to Unit-III, Page No. 81, Prob. 5

7. Two different types of drugs 'A' and 'B' were tried on certain patients for increasing weight. 6 persons were given drug 'A" and 8 persons were given drug 'B'. The increase in pounds is given below,

Drug 'A'	7	10	13	12	4	8		
Drug 'B'	12	8	3	18	16	9	8	3

Do the drugs 'A' and 'B' differ significantly with regard to their effect in increase in weight?

Sol : **(Imp.)**

Refer to Unit-III, Page No. 85, Prob. 7

UNIT - IV

1. Define Hypothesis. Explain the procedure for testing a Hypothesis.

Ans : **(Nov.-23, Imp.)**

Refer to Unit-IV, Page No. 90, Q.No. 1

2. How Hypothesis testing of mean can be done when the population standard deviation is them known with an examples.

Ans : **(Imp.)**

Refer to Unit-IV, Page No. 93, Q.No. 4

3. Write about Hypothesis testing - Difference of proportions.

Ans : **(Imp.)**

Refer to Unit-IV, Page No. 98, Q.No. 7

4. A manufacturer of pet animal foods was wondering whether cat owners and dog owners reacted differently to premium pet foods. They commissioned a consumer survey that yielded the following data.

Pet	Number of owners Surveyed	Number of owners using Premium food
Cat	280	152
Dog	190	81

is it reasonable to conclude at a = 0.05, the cat owners are more likely to feed their pets. Premium food than dog owners?

Sol : **(Imp.)**

Refer to Unit-IV, Page No. 100, Prob.7

5. Write about Hypothesis testing of two means.

Ans : **(Imp.)**

Refer to Unit-IV, Page No. 104, Q.No. 9

6. What are dependent samples? How to test hypothesis with dependent samples.

Ans : **(Imp.)**

Refer to Unit-IV, Page No. 109, Q.No. 11

7. Explain briefly about Hypothesis Concerning two Proportion.

Ans : **(Imp.)**

Refer to Unit-IV, Page No. 112, Q.No. 13

UNIT - V

1. Explain about Chi-square test.

Ans : **(Imp.)**

Refer to Unit-V, Page No. 117, Q.No. 1

2. Explain briefly about test for goodness of fit?

Ans : **(Imp.)**

Refer to Unit-V, Page No. 118, Q.No. 5

3. What is ANOVA? What are its assumptions and applications?

Ans : **(Imp.)**

Refer to Unit-V, Page No. 124, Q.No. 6

4. Explain briefly about F-Distribution.

Ans : **(Imp.)**

Refer to Unit-V, Page No. 139, Q.No. 9

5. Define Correlation. Explain different types of Correlation.

Ans : **(Imp.)**

Refer to Unit-V, Page No. 143, Q.No. 10

6. What is Karl Pearson's Coefficient of Correlation? Explain properties of Coefficient of Correlation.

Ans : **(Imp.)**

Refer to Unit-V, Page No. 145, Q.No. 13

8. Define Regression. State the uses of Regression.

Ans : **(Imp.)**

Refer to Unit-V, Page No. 150, Q.No. 15

9. What are the limitations of Regression Analysis?

Ans : **(Imp.)**

Refer to Unit-V, Page No. 151, Q.No. 18

10. How to find multiple regression equations.

Ans : **(Imp.)**

Refer to Unit-V, Page No. 159, Q.No. 24

11. Write about Inferences about individual slope Bi for population parameters.

Ans : **(Imp.)**

Refer to Unit-V, Page No. 162, Q.No. 25

UNIT I

Vector Spaces - Vector Spaces and Subspaces - Null Spaces, Column Spaces and Linear Transformations. Linearly Independent Sets - Bases - Coordinate Systems.

1.1 Vector Spaces and Subspaces

Q1. Explain about Vector Spaces and Subspaces.

Ans : **(Imp.)**

i) Vector Space

A vector space (V) is a non empty set containing objects or vectors (i.e., u, v, w) on which are defined two operations called addition and multiplication by scalars, that satisfy the following conditions,

i) $u + v \in V \; \forall \; u, v \in V$

ii) $u + v = v + u \; \forall \; u, v \in V$

iii) $(u + v) + = u + (v + w) \; \forall \; u, w \in V$

iv) Vector V contains a unique zero vector 0, such that, $u + 0 = u, \; \forall \; u \in V$.

v) Vector V contains a unique vector "–u" called negative inverse of u, such that, $u + (–u) = 0, \; \forall \; u \in V$.

vi) A scalar multiplication property is defined such that, if u is any vector in V and c and d are any two scalar, then,

- $c(du) = (cd)u$
- $(c + d)u = cu + du$
- $c(u + v) = cu + cv$
- $1u = u$

ii) Vector Subspace

A subset H of a vector space V is called subspace of V if it satisfy the following three conditions.

i) The zero vector of V is in H i.e., $0 \in H$.

ii) H is closed under vector addition i.e., for each $u, V \in H$, $u + v \in H$.

iii) H is closed under scalar multiplication i.e., For each $u \in H$, there exists a scalar 'c' such that $cu \in H$.

Problems

1. Let H be the set of all vector of the form $\begin{bmatrix} 2t \\ 0 \\ -t \end{bmatrix}$ **show that. H is subspace of R^3.**

Sol :

Given set H is set of all vectors of the form

$$\begin{bmatrix} 2t \\ 0 \\ -t \end{bmatrix}$$

Clearly the H = Span {v}

The vector set H can be written as

$$H = t \begin{bmatrix} 2 \\ 0 \\ -1 \end{bmatrix}$$

Which shows that span of {v}

$$\Rightarrow V = \begin{bmatrix} 2 \\ 0 \\ -1 \end{bmatrix}$$

$\therefore$ H is a subspace of R^3.

2. Let w be the set of all vectors of the form $\begin{bmatrix} 2b+3c \\ -b \\ 2c \end{bmatrix}$ where b & c are arbitrary. Find vector u & v such that w = span {u, v}. Does W is a subspace of R^3.

Sol :

Given set H is the set of all vectors of the form $\begin{bmatrix} 2b+3c \\ -b \\ 2c \end{bmatrix}$

Clearly the set H = Span {V} the vector set H can be written as $H = b\begin{bmatrix} 2 \\ -1 \\ 0 \end{bmatrix} + c\begin{bmatrix} 3 \\ 0 \\ 2 \end{bmatrix}$ which shows that spans of {V}

$$V = \begin{bmatrix} 2 \\ -1 \\ 0 \end{bmatrix} \quad u = \begin{bmatrix} 3 \\ 0 \\ 2 \end{bmatrix}$$

which means that energy element in the set can be written as a linear combination of the two vectors.

$$\therefore \quad H = \text{Span}\left\{\begin{bmatrix} 2 \\ -1 \\ 0 \end{bmatrix}\begin{bmatrix} 3 \\ 0 \\ 2 \end{bmatrix}\right\}$$

$\therefore$ H is a subspace of $R^3(R)$.

3. Verify the set W = {(x, y, 0)/x, y $\in$ F)} forms a subspace of $V_3(F)$.

Sol :

Given w = {x, y, 0)/x, y $\in$ F}

Let $\alpha = (x_1, y_1, 0)$ and $\beta = (x_2, y_2, 0)$ where $\alpha, \beta \in w$ and x, y $\in$ F.

Let a, b $\in$ F $\Rightarrow a\alpha + b\beta \in W$.

Now consider,

$$a\alpha + b\beta = a(x_1, y_1, 0) + b(x_2, y_2, 0)$$

$$a\alpha + b\beta = (ax_1, ay_1, 0) + (bx_2, by_2, 0)$$

$$a\alpha + b\beta = (ax_1 + bx_2, ay_1 + by_2, 0) \in W$$

$\therefore a\alpha + b\beta \in W \; \forall \alpha, \beta \in W$ and a, b $\in$ F

Here W is a subspace of V(F).

4. W{(x, y, z)/x – 3y + 4z = 0} show that w is a subspace of V(R).

Sol :

Given $W = \{(x, y, z)/x - 3y + 4z = 0\}$

Let $\alpha = (x_1, y_1, z_1)$ and $\beta = (x_2, y_2, z_2)$ and $a,b \in W$ and $a,b \in F$.

Now consider

$$a\alpha + b\beta = \alpha(x_1\ y_1\ z_1) + b(x_2, y_2, z_2)$$

$$\Rightarrow \quad a\alpha + b\beta = (\alpha x_1, ay_1, az_1) + (bx_2, by_2, bz_2)$$

Let $p = ax_1 + bx_2\ q = ay_1 + by_2 - r = az_1 + bz_2$

Now consider the given condition

$$P - 3q + 4\gamma = ax_1 + bx_2 - 3(ay_1 + by_2) + 4(az_1 + bz_2)$$

Since $\alpha = (x_1, y_1, z_1) \in W \Rightarrow x_1 - 3y_1 + 4z_1 = 0$ ... (1)

Since $\beta = (x_2, y_2, z_2) \in W \Rightarrow x_2 - 3y_2 + 4z_2 = 0$ ... (2)

$$p - 3\ q + 4\gamma = (ax_1 - 3ay_1 + 4az_1) + bx_2 - 3by_2 + 4bz_2)$$

$$= a(x_1 - 3y_1 + 4z_1) + b(x_2 - 3y_2 + 4z_2)$$

$$= a.0 + b.0$$

$\therefore P - 3q + 4r = 0\ \forall \alpha, \beta \in W$ and $a,b \in \mathbb{R}$.

5. Prove the set of solutions (x, y, z) of the equations x + y + 2z = 0 is a subspace of the space $R^3(R)$.

Sol :

Given $W = \{(x, y, z) / x + y + 2z = 0\}$ is a given set

Let $\alpha = (x_1, y_1, z_1)$ and $\beta = (x_2, y_2, z_2)$ where $\alpha, \beta \in W$

Let $a,b \in R$ and $\alpha, \beta \in W \Rightarrow a\alpha + b\beta \in W$

Consider $a\alpha + b\beta = a(x_1, y_1, z_1) + b(x_2, y_2, z_2)$

$$a\alpha + b\beta = (ax_1, ay_1, az_1) + (bx_2, by_2, bz_2)$$

$$a\alpha + b\beta = (ax_1 + bx_2, ay_1 + by_2, az_1 + bz_2)$$

Let us consider $p = ax_1 + bx_2, q = ay_1 + by_2, \gamma = az_1 + bz_2$

Since $\alpha \in W \Rightarrow x_1 + y_1 + 2z_1 = 0$ ... (1)

Since $\beta \in W \Rightarrow x_2 + y_2 + 2z_2 = 0$ ... (2)

$$p + q + 2\gamma = ax_1 + bx_2 + ay_1 + by_2 + 2(az_1 + bz_2)$$

$$p + q + 2\gamma = ax_1 + bx_2 + ay_1 + by_2 + 2az_1 + 2bz_2$$

$$= (ax_1 + ay_1 + 2az_1) + (bx_2 + by_2 + 2bz_2)$$

$$= a(x_1 + y_1 + 2z_1) + b(x_2 + y_2 + 2z_2)$$

$$= a.0 + b.0$$

$$p + q + 2\gamma = 0$$

$\therefore$ The above condition is satisfied.

$\therefore$ W is a subspace of $R^3(R)$.

6. V be vector space of 2×2 matrices over a field F. Show that W is not a subspace of V.

(i) $W = \{A \in V / \det A = 0\}$

(ii) $W = \{A \in V / A^2 = A\}$

Sol :

(i) $W = \{A \in V / \det A = 0\}$

To prove the above set let us consider

$$A = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \text{ and } B = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \text{ where } A, B \in W$$

Such that $\det A = 0$ & $\det B = 0$

In order to verify the above condition we verify whether $A, B \in W$ and $P, q \in F$

$\Rightarrow PA + Bq \in W$

$$\text{Consider } PA + Bq = p\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} + q\begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$$

$$PA + qB = \begin{bmatrix} p & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & q \end{bmatrix} = \begin{bmatrix} p & 0 \\ 0 & q \end{bmatrix}$$

$$\text{Now } PA + qB = \begin{bmatrix} p & 0 \\ 0 & q \end{bmatrix}$$

Now if we consider $|pA + qB| = pq \neq 0$

$\therefore$ W is not a subspace of V(F)

(ii) $W = \{A \in V / A^2 = A\}$

$$\text{Now we consider } A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$\Rightarrow A^2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = A$$

$\therefore A^2 = A$

$$\text{Similarly consider } B = \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix} \text{ such that}$$

$$B^2 = \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix}$$

$$\text{Now } A + B = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 2 & 1 \\ 0 & 1 \end{bmatrix}$$

Consider $(A+B)^2 = (A + B)(A + B)$

$$\Rightarrow \quad (A+B)^2 = \begin{bmatrix} 2 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 2 & 1 \\ 0 & 1 \end{bmatrix}$$

$$\Rightarrow \quad (A+B)^2 = \begin{bmatrix} 4 & 3 \\ 0 & 1 \end{bmatrix} \neq (A + B)$$

$\therefore (A + B)^2 \neq (A + B) \forall, B \in W$.

$\therefore$ W is not a subspace of V(F).

1.2 Null Spaces, Column Spaces and Linear Transformations

Q2. Define null space of an m × n matrix A.

Ans :

Null space of an m × n matrix A is the set of all solutions to the homogeneous equation.

$\therefore \quad Ax = 0$

Which is represented in set notation form as

Null A = {x : x is in R^n & A X = 0}

which is denoted by NU|A.

Problems

7. $\mathbf{A = \begin{bmatrix} 1 & -3 & -2 \\ -5 & 9 & 1 \end{bmatrix} \quad U = \begin{bmatrix} 5 \\ 3 \\ -2 \end{bmatrix}}$ **Determine if U belongs to the Null space of A.**

Sol : **(Imp.)**

Given $A = \begin{bmatrix} 1 & -3 & -2 \\ -5 & 9 & 1 \end{bmatrix}$ $U = \begin{bmatrix} 5 \\ 3 \\ -2 \end{bmatrix}$ if U is the Null space of A if it satifies AU = 0

$$AU = \begin{bmatrix} 1 & -3 & -2 \\ -5 & 9 & 1 \end{bmatrix} \begin{bmatrix} 5 \\ 3 \\ -2 \end{bmatrix}$$

$$= \begin{bmatrix} 5 & -9 & +4 \\ -25 & +27 & -2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Thus U is the Null space of A.

Q3. Prove that null space of m × n matrix A is a subspace of R^n.

Ans :

Clearly Nullspace of R^n has 'n' columns.

We know that nullspace A is said to be subsapce of R^n if

a) 0 is in Null space A

b) u + v is in null space A

c) CU is in Nullspace A

Now let the consider U, V are any two vector in the Null space A such that AU = 0 and AV = 0

Now to show that u + v is in the Nullspace A Consider,

$$A(U + V) = AU + AV$$
$$= 0 + 0$$
$$A(U + V) = 0$$

$\therefore$ U + V is in null space A

Let 'C' be the scalar such that

$$A(CU) = C(AU) = C(0) = 0$$

$\therefore$ A(CU) = 0

$\therefore$ CU is in Null space A

$\therefore$ Null space A is a a subsapce of R^n.

Q4. Define column space of a (m × n) matrix A .

Ans :

The column space of a m × n matrix A is the set of linear combinations of the columns of A

i.e., A = $\{a_1, a_2, a_3 \ldots a_n\}$ then

Column space A = Span $\{a_1, a_2, \ldots a_n\}$ which is denoted by ColA.

PROBLEMS

8. Find a matrix A such that

W = ColA where

$$\mathbf{W = \left\{ \begin{bmatrix} 6a-b \\ a+b \\ -7a \end{bmatrix} \right\}\ a, b \text{ in } R.}$$

Sol :

Given W = Col A

$$W = \left\{ \begin{bmatrix} 6a-b \\ a+b \\ -7a \end{bmatrix} : a, b \text{ in } R \right\}$$

Initillay writing was the linear combination.

$$W = \left\{ a\begin{bmatrix} 6 \\ 1 \\ -7 \end{bmatrix} + b\begin{bmatrix} -1 \\ 1 \\ 0 \end{bmatrix} : a, b \in R \right\}$$

$$W = \text{Span} \left\{ \begin{bmatrix} 6 \\ 1 \\ -7 \end{bmatrix} \begin{bmatrix} -1 \\ 1 \\ 0 \end{bmatrix} \right\}$$

Now the vectors in the spanning set A as the columns of A which is $A = \begin{bmatrix} 6 & -1 \\ 1 & 1 \\ -7 & 0 \end{bmatrix}$

9. Let $A = \begin{bmatrix} 2 & 4 & -2 & 1 \\ -2 & -5 & 7 & 3 \\ 3 & 7 & -8 & 6 \end{bmatrix}$

a) If the column space of A is a subspace of R^k, what is k?

b) If the Null space of A is a subspace of R^k what is k?

Sol :

Given matrix is

$$A = \begin{bmatrix} 2 & 4 & -2 & 1 \\ -2 & -5 & 7 & 3 \\ 3 & 7 & -8 & 6 \end{bmatrix}$$

Here the matrix A is 3 × 4 matrix

a) The columns of A each have three entries, So colA is a subspace of R^k. where k = 3.

b) Similarly the vector X such that A x is defined must have four entries so NulA is a subspace of R^k where K = 4.

10. Determine if $W = \begin{bmatrix} 1 \\ 3 \\ -4 \end{bmatrix}$ is in Null A where $A = \begin{bmatrix} 3 & -5 & -3 \\ 6 & -2 & 0 \\ -8 & 4 & 1 \end{bmatrix}$

Sol :

By def of NU|A the have AW = 0

Now consider

$$\begin{bmatrix} 3 & -5 & -3 \\ 6 & -2 & 0 \\ -8 & 4 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 3 \\ 4 \end{bmatrix}$$

$$= \begin{bmatrix} 3 & +3(-5) & -4(-3) \\ 1.6 & +3(-2) & -4(0) \\ 1(-8) & +3.4 & -4.1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

$\therefore$ W is in the Null space of A.

---

11. Determine if $W = \begin{bmatrix} 1 \\ -1 \\ 1 \end{bmatrix}$ is in Nul A where $A = \begin{bmatrix} 2 & 6 & 4 \\ -3 & 2 & 5 \\ -5 & -4 & 1 \end{bmatrix}$

Sol :

Given $A = \begin{bmatrix} 2 & 6 & 4 \\ -3 & 2 & 5 \\ -5 & -4 & 1 \end{bmatrix}$

by def of NUIA we have AW = 0 Now consider

$$\begin{bmatrix} 2 & 6 & 4 \\ -3 & 2 & 5 \\ -5 & -4 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ -1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1.(2) & -1.(6) & +1.4 \\ 1(-3) & -1(2) & +1.5 \\ 1(-5) & -1(-4) & +1.1 \end{bmatrix}$$

$$= \begin{bmatrix} 2 & -6 & +4 \\ -3 & -2 & +5 \\ -5 & +4 & +1 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

$\therefore$ W is in the Null space of A.

---

12. For the matrix $A = \begin{bmatrix} 4 & 5 & -2 & 6 & 0 \\ 1 & 1 & 0 & 1 & 0 \end{bmatrix}$

(i) Find k such that Nul A is a subspace of R^K.

(ii) Find k such that Col A is a subspace of R^k.

Sol :

Given matrix is $A = \begin{bmatrix} 4 & 5 & -2 & 6 & 0 \\ 1 & 1 & 0 & 1 & 0 \end{bmatrix}$

Here A is a 2 × 5 matrix

a) Nul A is subspace of R^n which is R^5

$\therefore \quad K = 5$

b) Col A is subspace of R^m(ie) R^2.

$\therefore \quad K = 2$

---

13. For the matrix A = $\begin{bmatrix} 6 & -4 \\ -3 & 2 \\ -9 & 6 \\ 9 & -6 \end{bmatrix}$ Find the K such that Nul A is a subspace of R^k. Find K such that Col A is a subspace of R^k.

Sol :

Given matrix is

$$A = \begin{bmatrix} 6 & -4 \\ -3 & 2 \\ -9 & 6 \\ 9 & -6 \end{bmatrix}$$

Here A is 4 × 2 matrix

a) Nul A is subspace of R^n which is R^2.

$\therefore \quad K = 2$

b) Col A is subspace of R^m is R^4

$\therefore \quad K = 4$

---

14. Find the matrix A such that the set

S = $\left\{ \begin{bmatrix} b-c \\ 2b+c+d \\ 5c-4d \\ d \end{bmatrix} \right\}$: b, c, d real is col A.

Sol :

Given H = $\left\{ \begin{bmatrix} b-c \\ 2b+c+d \\ 5c-4d \\ d \end{bmatrix} \right\}$

Where the set can be written as

$$\left\{ \begin{bmatrix} b-c \\ 2b+c+d \\ 5c-4d \\ d \end{bmatrix} \right\} = b \begin{bmatrix} 1 \\ 2 \\ 0 \\ 0 \end{bmatrix} + c \begin{bmatrix} -1 \\ 1 \\ 5 \\ 0 \end{bmatrix} + d \begin{bmatrix} 0 \\ 1 \\ -4 \\ 1 \end{bmatrix}$$

$$= \text{span} \left\{ \begin{bmatrix} 1 \\ 2 \\ 0 \\ 0 \end{bmatrix} \begin{bmatrix} -1 \\ 1 \\ 5 \\ 0 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ -4 \\ 1 \end{bmatrix} \right\}$$

Since by hypothesis H = Col A

$$\therefore \quad A = \begin{bmatrix} 1 & -1 & 0 \\ 2 & 1 & 1 \\ 0 & 5 & -4 \\ 0 & 0 & 1 \end{bmatrix}$$

---

15. Find A such that the set $\left\{ \begin{bmatrix} b-c \\ 2b+3d \\ b+3c-3d \\ c+d \end{bmatrix} \right\}$ is Col A **(Imp.)**

Sol :

Let H = $\left\{ \begin{bmatrix} b-c \\ 2b+3d \\ b+3c-3d \\ c+d \end{bmatrix} \right\}$

The set H can be written as linear combination of vectors.

$$H = b \begin{bmatrix} 1 \\ 2 \\ 1 \\ 0 \end{bmatrix} + c \begin{bmatrix} -1 \\ 0 \\ 3 \\ 1 \end{bmatrix} + d \begin{bmatrix} 0 \\ 3 \\ -3 \\ 1 \end{bmatrix}$$

$$= \text{span} \left\{ \begin{bmatrix} 1 \\ 2 \\ 1 \\ 0 \end{bmatrix} \begin{bmatrix} -1 \\ 0 \\ 3 \\ 1 \end{bmatrix} \begin{bmatrix} 0 \\ 3 \\ -3 \\ 1 \end{bmatrix} \right\}$$

Since by hypothesis we have that H = Col A

16. Let $A = \begin{bmatrix} -2 & 4 \\ -1 & 2 \end{bmatrix}$ $W = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$ Determine if W is in Col A is W in Nul A?

Sol :

Given matrix,

$$A = \begin{bmatrix} -2 & 4 \\ -1 & 2 \end{bmatrix} \quad W = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$$

Col A is given by $AX = b \Rightarrow AX = W$

The Augmented matrix [AW] is

$$[AW] = \begin{bmatrix} -2 & 4 & 2 \\ -1 & 2 & 1 \end{bmatrix} R_2 \rightarrow 2R_2 - R_1$$

$$[AW] = \begin{bmatrix} -2 & 4 & 2 \\ 0 & 0 & 0 \end{bmatrix} R_1 \rightarrow R_1/-2$$

$$[AW] = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \end{bmatrix}$$

The system of equation AX = W is consistent

$\therefore$ W is in Col A

And Nul A is given by AX = 0

Consider

$$AW = \begin{bmatrix} -2 & 4 \\ -1 & 2 \end{bmatrix} \begin{bmatrix} 2 \\ 1 \end{bmatrix}$$

$$= \begin{bmatrix} -4 & +4 \\ -2 & +2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$\therefore$ AW = 0

$\therefore$ W is the solution of AW = 0

$\therefore$ W is in Nul A

Q5. Define linear transformation.

Sol :

If V and W represents two vector spaces than a mapping T: $V \rightarrow W$ is said to be linear transformation if the following conditions are satisfied.

(i) $T(u+v) = T(u) + T(v) \forall\ u, v \in V$

(ii) $T(cu) = cT(u) \forall\ u \in V$ and all scalars c.

Q6. Define Kernel and Range of a linear transformation.

Sol :

Kernel of a Linear Transformation

If V and W are two vector spaces and T is a linear transformation from V into W then the kernel or null space of a linear transformation N(T) is defined as the set of all vectors u in V such that, T(u) = 0 where, 0 is the zero vector in W.

It is represented in set notation form as,

$$N(T) = \{v \in V : T(u) = 0 \in W\}$$

Range of a Linear Transformation

If V and W are two vector spaces and T is a linear transformation from V into W then range of linear transformation R(T) is defined as the set of all vector in W of the form T(x) for some $x \in V$.

It is represented in set notation form as,

$$R(T) = \{T(x) : x \in V\}.$$

PROBLEMS

17. Define $T : P_2 \rightarrow R^2$ by $T(p) = \begin{bmatrix} p(0) \\ p(1) \end{bmatrix}$. For instance, if $p(t) = 3 + 5t + 7t^2$, then $T(p) = \begin{bmatrix} 3 \\ 15 \end{bmatrix}$. Show that T is a linear transformation.

Sol :

Given, $T : p_2 \rightarrow R^2$

$$\text{And } T(p) = \begin{bmatrix} p(0) \\ p(1) \end{bmatrix} \quad \ldots (1)$$

Let p, q be polynomials in P_2 i.e.,

$p, q \in P_2$

$$\text{Then, } T(p+q) = \begin{bmatrix} (p+q)(0) \\ (p+q)(1) \end{bmatrix}$$

$$= \begin{bmatrix} p(0)+q(0) \\ p(1)+q(1) \end{bmatrix}$$

$$= \begin{bmatrix} p(0) \\ p(1) \end{bmatrix} + \begin{bmatrix} q(0) \\ q(1) \end{bmatrix}$$

$$= T(p) + T(q)$$

[$\because$ From equation (1)]

$\therefore$ $T(p+q) = T(p) + T(q)$

Let c be any scalar,

Then $T(cp) = \begin{bmatrix} (cp)(0) \\ (cp)(1) \end{bmatrix}$

$$= \begin{bmatrix} cp(0) \\ cp(1) \end{bmatrix}$$

$$= c\begin{bmatrix} p(0) \\ p(1) \end{bmatrix}$$

$$= cT(p)$$

$\therefore$ $T(cp) = cT(p)$

$\therefore$ T is a linear transformation.

18. Define a linear transformation $T : P_2 \rightarrow R^2$ by $T(p) = \begin{bmatrix} p(0) \\ p(0) \end{bmatrix}$. For polynomials P_1 and P_2 in P_2 that span the kernel of T, and describe the range of T.

Sol :

Given,

$T : P_2 \rightarrow R^2$ is a linear transformation and,

$$T(p) = \begin{bmatrix} p(0) \\ p(0) \end{bmatrix}$$

Let q be any quadratic polynomial such that it will be in kernel of T i.e.,

$q(0) = 0$

Then, $q = at + bt^2$

$\therefore$ The polynomial $p_1(t) = t$ and $p_2(t) = t^2$ which span kernel of T.

Let the range of T be of the form $\begin{bmatrix} a \\ a \end{bmatrix}$ and p(t) a.

Here the vector is the image of polynomial.

$\therefore$ Range of $T = \left\{ \begin{bmatrix} a \\ a \end{bmatrix} \forall a \in R \right\}$.

19. Let $M_{2\times2}$ be the vector space of all 2×2 matrices, and define T: $M_{2\times2} \rightarrow M_{2\times2}$ by $T(A) = A + A^T$, where $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$.

(i) Show that T is a linear transformation.

(ii) Let B be any element of $M_{2\times2}$ such that $B^T = B$. Find an A in $M_{2\times2}$ such that T(A) = B.

(iii) Show that the range of T is the set of B in $M_{2\times2}$ with the property that $B^T = B$.

(iv) Describe the kernel of T.

Sol :

Given,

$M_{2\times2}$ is a vector space.

And $T : M_{2\times2} \rightarrow M_{2\times2}$ is defined by,

$T(A) = A + A^T$ ... (1)

Here $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$

(i) Let $A, B \in M_{2\times2}$

Consider,

$T(A + B) = (A + B) + (A + B)^T$

[$\therefore$ From equation (1)]

$= A + B + A^T + B^T$

$= (A + A^T) + (B + B^T)$

$= T(A) + T(B)$

$\therefore$ $T(A + B) = T(A) + T(B)$

Let there exist a scalar c such that,

$T(cA) = (cA) + (cA)^T$

$= cA + cA^T$

[$\therefore$ From equation (1)]

$= c(A + A^T)$

$\therefore$ $T(cA) = cT(A)$

$\therefore$ T is a linear transformation.

(ii) Given $B \in M_{2\times 2}$ such that $B^T = B$.

Let, $A = \frac{1}{2}B$

Consider, $T(A) = A + A^T$

$= \frac{1}{2}B + \left(\frac{1}{2}B\right)^T$

$= \frac{1}{2}B + \frac{1}{2}B^T$

$= \frac{1}{2}B + \frac{1}{2}B$ [$\therefore B^T = B$]

$= B$

$\therefore$ $T(A) = B$

(iii) Let B be in range of T,

Then $B = T(A)$

$B = A + A^T$ [$\therefore$ From equation (1)]

And $B^T = (A + A^T)^T$

$= A^T + (A^T)^T$

$= A^T + A$

$= B$

$\therefore$ $B^T = B$

$\therefore$ B has the property $B^T = B$.

(iv) Let $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$ be in kernel of T,

i.e., $T(A) = A + A^T = 0$

Consider,

$$A + A^T = \begin{bmatrix} a & b \\ c & d \end{bmatrix} + \begin{bmatrix} a & c \\ b & d \end{bmatrix}$$

$$= \begin{bmatrix} 2a & b+c \\ b+c & 2d \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$$

On comparing both sides,

$2a = 0 \quad \Rightarrow a = 0$

$b + c = 0 \quad \Rightarrow c = -b$

$2d = 0 \quad \Rightarrow d = 0$

$\therefore$ Kernel of T is, $\left\{ \begin{bmatrix} 0 & b \\ -b & 0 \end{bmatrix} \forall b \in R \right\}$

1.3 Linearly Independent Sets, Bases

Problems

20. Determine which of the following sets are bases for R^3. Of the sets that are not bases, determine which ones are linearly independent and which ones span R^3.

(i) $\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$

(ii) $\begin{bmatrix} 2 \\ -2 \\ 1 \end{bmatrix} \begin{bmatrix} 1 \\ -3 \\ 2 \end{bmatrix} \begin{bmatrix} -7 \\ 5 \\ 4 \end{bmatrix}$

(iii) $\begin{bmatrix} 1 \\ -3 \\ 0 \end{bmatrix} \begin{bmatrix} -2 \\ 9 \\ 0 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \begin{bmatrix} 0 \\ -3 \\ 5 \end{bmatrix}$

(iv) $\begin{bmatrix} -2 \\ 3 \\ 0 \end{bmatrix} \begin{bmatrix} 6 \\ -1 \\ 5 \end{bmatrix}$

Sol :

(i) Given set of vectors,

$$\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

Let, $A = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix}$

$R_1 \rightarrow R_1 - R_2$

$$A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix}$$

$R_2 \rightarrow R_2 - R_3$

$$A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

The matrix A has 3 pivot positions,

$\therefore$ From invertible matrix theorem,

A is invertible and forms a basis for R^3.

(ii) Given set of vectors,

$$\begin{bmatrix} 2 \\ -2 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ -3 \\ 2 \end{bmatrix}, \begin{bmatrix} -7 \\ 5 \\ 4 \end{bmatrix}$$

Let $A = \begin{bmatrix} 2 & 1 & -7 \\ -2 & -3 & 5 \\ 1 & 2 & 4 \end{bmatrix}$

$R_2 \rightarrow R_2 + R_1, R_3 \rightarrow 2R_3 - R1$

$$A = \begin{bmatrix} 2 & 1 & -7 \\ 0 & -2 & -2 \\ 0 & 3 & 1 \end{bmatrix}$$

$R_2 \rightarrow 2R_3 + 3R_2$

$$A = \begin{bmatrix} 2 & 1 & -7 \\ 0 & -2 & -2 \\ 0 & 0 & -4 \end{bmatrix}$$

$R_1 \rightarrow 2R_1 + R_2, R_2 \rightarrow \frac{R_2}{-2}, R_3 \rightarrow \frac{R_3}{-4}$

$$A = \begin{bmatrix} 4 & 0 & -16 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix}$$

$R_2 \rightarrow R_2 - R_3$

$$A = \begin{bmatrix} 4 & 0 & -16 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$R_1 \rightarrow \frac{R_1}{4}$

$$A = \begin{bmatrix} 1 & 0 & -4 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$R_1 \rightarrow R_1 + 4R_3$

$$A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

The matrix A has 3 pivot positions,

$\therefore$ From invertible matrix theorem

A is invertible and forms a basis for R^3.

(iii) Given set of vectors,

$$\begin{bmatrix} 1 \\ -3 \\ 0 \end{bmatrix} \begin{bmatrix} -2 \\ 9 \\ 0 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \begin{bmatrix} 0 \\ -3 \\ 5 \end{bmatrix}$$

Since, the set contains zero vector,

$\Rightarrow$ Set is not linearly independent

$$\text{Let, } A = \begin{bmatrix} 1 & -2 & 0 & 0 \\ -3 & 9 & 0 & -3 \\ 0 & 0 & 0 & 5 \end{bmatrix}$$

$R_2 \to R_2 + 3R_1$

$$A = \begin{bmatrix} 1 & -2 & 0 & 0 \\ 0 & 3 & 0 & -3 \\ 0 & 0 & 0 & 5 \end{bmatrix}$$

$R_2 \to \frac{R_2}{3}, R_3 \to \frac{R_3}{5}$

$$A = \begin{bmatrix} 1 & -2 & 0 & 0 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$R_1 \to R_1 + 2R_2$

$$A = \begin{bmatrix} 1 & 0 & 0 & -2 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$R_1 \to R_1 + 2R_3$

$R_2 \to R_2 + R_3$

$$A = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

The matrix A has a pivot in each row.

$\therefore$ The given vectors spans R^3.

(iv) Given set of vectors,

$$\begin{bmatrix} -2 \\ 3 \\ 0 \end{bmatrix} \begin{bmatrix} 6 \\ -1 \\ 5 \end{bmatrix}$$

$$\text{Let, } A = \begin{bmatrix} -2 & 6 \\ 3 & -1 \\ 0 & 5 \end{bmatrix}$$

$R_2 \to 2R_2 + 3R_1$

$$A = \begin{bmatrix} -2 & 6 \\ 0 & 16 \\ 0 & 5 \end{bmatrix}$$

$R_1 \to \frac{R_1}{-2}$

$R_2 \to \frac{R_2}{16}$

$$A = \begin{bmatrix} 1 & -3 \\ 0 & 1 \\ 0 & 5 \end{bmatrix}$$

$R_1 \to R_1 + 3R_2$

$R_3 \to R_3 - 5R_2$

$$A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 5 \end{bmatrix}$$

The matrix has pivot in each column

$\Rightarrow$ Given vectors are linearly independent

Also it can be seen that, the matrix contains no pivot in each row.

$\therefore$ Given vectors does not span R^3.

21. Find bases for the null spaces of the matrix.

$$\begin{bmatrix} \mathbf{1} & \mathbf{0} & \mathbf{-3} & \mathbf{2} \\ \mathbf{0} & \mathbf{1} & \mathbf{-5} & \mathbf{4} \\ \mathbf{3} & \mathbf{-2} & \mathbf{1} & \mathbf{-2} \end{bmatrix}.$$

Sol :

Given matrix is,

$$\begin{bmatrix} 1 & 0 & -3 & 2 \\ 0 & 1 & -5 & 4 \\ 3 & -2 & 1 & -2 \end{bmatrix}$$

$R_3 \rightarrow R_3 - 3R_1$

$$= \begin{bmatrix} 1 & 0 & -3 & 2 \\ 0 & 1 & -5 & 4 \\ 0 & -2 & 10 & -8 \end{bmatrix}$$

$R_3 \rightarrow \frac{R_3}{2}$

$$= \begin{bmatrix} 1 & 0 & -3 & 2 \\ 0 & 1 & -5 & 4 \\ 0 & -1 & 5 & -4 \end{bmatrix}$$

$R_3 \rightarrow R_3 + 3R_2$

$$= \begin{bmatrix} 1 & 0 & -3 & 2 \\ 0 & 1 & -5 & 4 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

The general solution is,

$x_1 - 3x_3 + 2x_4 = 0$

$\Rightarrow \quad x_1 = 3x_3 - 2x_4$

$x_2 - 5x_3 + 4x_4 = 0$

$\Rightarrow \quad x_3 = 5x_3 - 4x_4$

Here x_3 and x_4 are free variables

$$\therefore \quad X = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 3x_3 - 2x4 \\ 5x_3 - 4x_4 \\ x_3 \\ x_4 \end{bmatrix}$$

$$X = x_3 \begin{bmatrix} 3 \\ 5 \\ 1 \\ 0 \end{bmatrix} + x_4 \begin{bmatrix} -2 \\ -4 \\ 0 \\ 1 \end{bmatrix}$$

$\therefore$ The basis for Nul A is,

$$\left\{ \begin{bmatrix} 3 \\ 5 \\ 1 \\ 0 \end{bmatrix} \begin{bmatrix} -2 \\ -4 \\ 0 \\ 1 \end{bmatrix} \right\}$$

22. Find a basis for the set of vectors in R^3 in the plane x + 2y + z = 0.

Sol : **(Imp.)**

Given plane is,

$x + 2y + z = 0$ ... (1)

Let, A = [1 2 1]

From equation (1) the general solution is,

$x = -2y - z$

Let y and z be free variables

$$\therefore \; X = \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} -2y - z \\ y \\ z \end{bmatrix}$$

$$= y \begin{bmatrix} -2 \\ 1 \\ 0 \end{bmatrix} + z \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix}$$

$\therefore$ The basis for Nul A is,

$$\left\{ \begin{bmatrix} -2 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix} \right\}$$

23. Determine whether the following set of vectors {(1, –2, 1), (2, 1, –1), (7, – 4, 1)} is linearly dependent or linearly independent.

Sol :

Given set of vectors,

$\{(1, -2, 1), (2, 1, -1), (7, -4, 1)\}$

Let, $v_1 = \begin{bmatrix} 0 \\ -2 \\ 0 \end{bmatrix}$, $v_2 = \begin{bmatrix} 2 \\ 1 \\ -1 \end{bmatrix}$, $v_3 = \begin{bmatrix} 7 \\ -4 \\ 1 \end{bmatrix}$

Consider the matrix,

$A = [v_1, v_2, v_3]$

$$= \begin{bmatrix} 1 & 2 & 7 \\ -2 & 1 & -4 \\ 1 & -1 & 1 \end{bmatrix}$$

$R_2 \to R_2 + 2R_1$

$R_3 \to R_3 - R_1$

$$= \begin{bmatrix} 1 & 2 & 7 \\ 0 & 5 & 10 \\ 0 & -3 & -6 \end{bmatrix}$$

$R_3 \to \frac{R_2}{5}$, $R_3 \to \frac{R_3}{-3}$

$$= \begin{bmatrix} 1 & 2 & 7 \\ 0 & 1 & 2 \\ 0 & 1 & 2 \end{bmatrix}$$

$R_3 \to R_3 - R_2$

$$= \begin{bmatrix} 1 & 2 & 7 \\ 0 & 1 & 2 \\ 0 & 1 & 0 \end{bmatrix}$$

$R_1 \to R_1 - 2R_2$

$$= \begin{bmatrix} 1 & 0 & 3 \\ 0 & 1 & 2 \\ 0 & 0 & 0 \end{bmatrix}$$

Since the matrix A does not contain pivot in each column.

$\Rightarrow$ The set is not linearly independent.

$\therefore$ The given set of vectors are linearly dependent.

24. Determine whether the set of vectors {(1, 1, 2), (1, 2, 5), (5, 3, 4)} forms a basis of R^3.

Sol :

Given set of vectors is,

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

Let $v_1 = \begin{bmatrix} 1 \\ 1 \\ 2 \end{bmatrix}$, $v_2 = \begin{bmatrix} 1 \\ 2 \\ 5 \end{bmatrix}$, $v_3 = \begin{bmatrix} 5 \\ 3 \\ 4 \end{bmatrix}$

Consider the matrix,

$$A = [v_1\ v_2\ v_3] = \begin{bmatrix} 1 & 1 & 5 \\ 1 & 2 & 3 \\ 2 & 5 & 4 \end{bmatrix}$$

$R_2 \to R_2 - R_1$

$R_3 \to R_3 - 2R_1$

$$= \begin{bmatrix} 1 & 1 & 5 \\ 1 & 2 & 3 \\ 2 & 5 & 4 \end{bmatrix}$$

$R_3 \to \frac{R_3}{-3}$

$$= \begin{bmatrix} 1 & 1 & 5 \\ 0 & 1 & -2 \\ 0 & -1 & 2 \end{bmatrix}$$

$R_3 \to R_3 + R_1$

$$= \begin{bmatrix} 1 & 1 & 5 \\ 0 & 1 & -2 \\ 0 & 0 & 0 \end{bmatrix}$$

$$R_1 \to R_1 - R_2$$

$$A = \begin{bmatrix} 1 & 0 & 7 \\ 0 & 1 & -2 \\ 0 & 0 & 0 \end{bmatrix}$$

Since the matrix A has only two pivot positions.

Thus the columns do not form a basis for R^3.

1.4 COORDINATE SYSTEMS

Q7. Define Coordinates?

Sol :

Let $S = \{\alpha_1, \alpha_2, ... \alpha_n\}$ be a basis set of a finite dimensional vector space V(F). Let $\beta \in V$ $\beta = a_1\alpha_1 + a_2\alpha_2 + ... + a_n\alpha_n \ \forall a_1, a_2, ... a_n \in F$ then $\exists$ scalars $(a_1, a_2...., a_n)$ are called coordiantes.

PROBLEMS

25. Determine whether set of vector S = {(1, 2, 1) (2, 1, 0) (1, –1, 2)}forms a basis of V_3 (F).

Sol :

Given set of vectors is

$$S = \{(1,2,1)(2,1,0)(1,-1,2)\}$$

Let $V_1 = \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix} V_2 = \begin{bmatrix} 2 \\ 1 \\ 0 \end{bmatrix} V_3 = \begin{bmatrix} 1 \\ -1 \\ 2 \end{bmatrix}$

Consider the matrix

$$A = [V_1 \quad V_2 \quad V_3] V_2 = \begin{bmatrix} 1 & 2 & 1 \\ 2 & 1 & -1 \\ 1 & 0 & 2 \end{bmatrix}$$

$$R_2 \to R_2 = 2$$

$$R_3 \to R_3$$

$$= \begin{vmatrix} 1 & 2 & 1 \\ 0 & -3 & -3 \\ 0 & -2 & 1 \end{vmatrix}$$

$$R_3 \to 3R_3 - 2R_2$$

$$= \begin{vmatrix} 1 & 2 & 1 \\ 0 & -3 & -3 \\ 0 & 0 & 9 \end{vmatrix}$$

$$R_2 \to \frac{R_2}{-3}$$

$$\begin{vmatrix} 1 & 2 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 9 \end{vmatrix}$$

$$R_3 \to \frac{R_3}{9}$$

$$\begin{vmatrix} 1 & 2 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{vmatrix}$$

$$R_2 \to R_2 - R_3$$

$$\begin{vmatrix} 1 & 2 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{vmatrix}$$

$$R_1 \to -2R_2 + R_1$$

$$\begin{vmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{vmatrix}$$

$$R_1 \to R_1 \to R_3$$

$$\begin{vmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{vmatrix}$$

Since the matrix A has only three point elements.

$\therefore$ S forms a basic set.

26. Show that the set $S = \{V_1\ V_2\ V_3\}$

$$V_1 \begin{bmatrix} 2 \\ 1 \\ 4 \end{bmatrix} V_2 = \begin{bmatrix} 1 \\ -1 \\ 2 \end{bmatrix} V_3 = \begin{bmatrix} 3 \\ 1 \\ -2 \end{bmatrix}.$$

from a basis for R^3.

Sol :

Given $\{V_1V_2V_3\}$

$$A = \begin{bmatrix} 2 & 1 & 3 \\ 1 & -1 & 1 \\ 4 & 2 & -2 \end{bmatrix}$$

$R_1 \leftrightarrow R_2$

$$\begin{bmatrix} 1 & -1 & 1 \\ 2 & 1 & 3 \\ 4 & 2 & -2 \end{bmatrix}$$

$R_2 \rightarrow R_2 - 2R_1$

$R_3 \rightarrow R_3 - 4R_1$

$$\begin{bmatrix} 1 & -1 & 1 \\ 0 & 3 & 1 \\ 0 & 6 & -6 \end{bmatrix}$$

$R_3 \rightarrow R_3 - 2R_2$

$$\begin{bmatrix} 1 & -1 & 1 \\ 0 & 3 & 1 \\ 0 & 0 & -8 \end{bmatrix}$$

$R_3 \rightarrow \frac{R_3}{-8}$

$$\begin{bmatrix} 1 & -1 & 1 \\ 0 & 3 & 1 \\ 0 & 0 & 1 \end{bmatrix}$$

$R_3 \rightarrow R_2 - R_3$

$$\begin{bmatrix} 1 & -1 & 1 \\ 0 & 3 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$R_2 \rightarrow \frac{R_2}{3}$

$$\begin{bmatrix} 1 & -1 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$R_1 \rightarrow R_1 + R_2$

$$\begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$R_1 \rightarrow R_2 - R_3$

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Since the matrix A has three pivot elements thus the matrix A form a basis of P^2.

---

27. Let $V_1 = \begin{bmatrix} 1 \\ -2 \\ 3 \end{bmatrix} V_2 = \begin{bmatrix} -2 \\ +7 \\ -9 \end{bmatrix}$ Now determine of $\{V_1V_2\}$ is a basis for R^2.

Sol :

Given $V_1 = \begin{bmatrix} 1 \\ -2 \\ 3 \end{bmatrix} V_2 = \begin{bmatrix} -2 \\ 7 \\ -9 \end{bmatrix}$

Now to verify the set $S = \{V_1V_2\}$ is a basis we have to show that S is α.I and linear span of V.

Consider the matrix

$$A = \begin{bmatrix} 1 & -2 \\ -2 & 7 \\ 3 & -9 \end{bmatrix}$$

$R_2 \rightarrow R_2 + 2R_1$

$R_3 \rightarrow R_3 - 3R_1$

$$= \begin{vmatrix} 1 & -2 \\ 0 & 3 \\ 0 & -3 \end{vmatrix}$$

$R_3 \rightarrow R_3 + R_2$

$$\begin{vmatrix} 1 & -2 \\ 0 & 3 \\ 0 & 0 \end{vmatrix}$$

$R_2 \to R_2 / 3$

$$\begin{vmatrix} 1 & -2 \\ 0 & 1 \\ 0 & 0 \end{vmatrix}$$

$R_1 \to R_1 + R_2$

$$\begin{bmatrix} 1 & -1 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}$$

$R_2 \to R_2 + R_1$

$$\begin{vmatrix} 1 & -1 \\ 0 & 1 \\ 0 & 0 \end{vmatrix}$$

$R_1 \to R_1 + R_2$

$\therefore$ S is basis of V.

28. Suppose $\{\alpha_1\alpha_2,.....\alpha_K\}$ is L.D spanning set for a vector space V. Show that each W in V can be expressed in more then one way as a linear combination of $\alpha_1,\alpha_2.....\alpha_k$.

Sol :

Let $\{\alpha_1,\alpha_2,.....\alpha_k\}$ is Linearly dependent.

Then by def $\exists$ $c_1,c_2....c_k$ such that $c_1\alpha_1 + c_2\alpha_2 + ... + c_k\alpha_k = 0$ .... (1)

Let $W \in V\exists$ scales $K_1,K_2,.....K_K$ such that W is a linear combination of vector $\alpha_1,\alpha_2.....\alpha_k$.

$$W = K_1\alpha_1 + K_2\alpha_2 + ... + K_k\alpha_K \quad ....(2)$$

Now by adding the equations (1) & (2)

$$O + W = C_1\alpha_1 + C_2\alpha_2 + ... + C_k\alpha_k + K_1\alpha_1 + K_2\alpha_2 + ... + K_k\alpha_k$$

$$W = (C_1 + K_1)\alpha_1 + (C_2 + K_2)\alpha_2 + ... + (C_K + K_K)2_k$$

$\therefore$ W is expressed on the linear.

Combination of $\alpha_1,\alpha_2,\alpha_3.....\alpha_k$.

29. Use coordinate vectors to test whether the following set of polynomial span P^2. $1–3t+5t^2$, $–3 + 5t – 7t^2$, $– 4 +5t-6t^2$, and $1-t^2$.

Sol :

Given polynomial is

$1–3t+5t^2$, $–3 + 5t – 7t^2$, $– 4 +5t-6t^2$, & $1-t^2$.

$$\text{Let } A = \begin{bmatrix} 1 & -3 & -4 & 1 \\ -3 & 5 & 5 & 0 \\ 5 & -7 & -6 & -1 \end{bmatrix}$$

Now converting in to Echolon form

$R_2 \to R_2 + 3R_1$

$R_3 \to R_3 - 5R_1$

$$A = \begin{bmatrix} 1 & -3 & -4 & 1 \\ 0 & -4 & -7 & 3 \\ 0 & 8 & 14 & -6 \end{bmatrix}$$

$R_3 \to \frac{R_3}{2}$

$$A = \begin{bmatrix} 1 & -3 & -4 & 1 \\ 0 & -4 & -7 & 3 \\ 0 & 4 & 7 & -3 \end{bmatrix}$$

$R_3 \to R_3 + R_2$

$$A = \begin{bmatrix} 1 & -3 & -4 & 1 \\ 0 & -4 & -7 & 3 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

The matrix A does not contain a point in each row.

The vectors do not span R^3 & from isomophism of R^3 & R^2.

The given polynomials do not span P_2 .

30. Determine whether set of polynomial form a basis for P_3.

$5 - 3t + 4t^2 + 2t^3, 9 + t + 8t^2 - 6t^3, 6 - 2t + 5t^2, t^3$.

Sol :

Given set of polynomials are above then ∴ coordinate vector are (5,–3,4, 2) (9,1,8,–6) (6,–2,5,0) & (0,0,0,1)

$$A = \begin{bmatrix} 5 & 9 & 6 & 0 \\ -3 & 1 & -2 & 0 \\ 4 & 8 & 5 & 0 \end{bmatrix}$$

$R_2 \to 5R_2 + 3R_1$

$R_3 \to 5R_3 - 4R_1$

$R_4 \to 5R_4 - 2R_1$

$$A = \begin{bmatrix} 5 & 9 & 6 & 0 \\ 0 & 32 & 8 & 0 \\ 0 & 4 & 1 & 0 \\ 0 & -48 & -12 & -5 \end{bmatrix}$$

$$R_2 \to \frac{R_2}{8}, R_4 \to \frac{R_4}{12}$$

$$A = \begin{vmatrix} 5 & 9 & 6 & 0 \\ 0 & 4 & 1 & 0 \\ 0 & 4 & 1 & 0 \\ 0 & -4 & -1 & \frac{5}{12} \end{vmatrix}$$

$$R_3 \to R_3 + R_2$$

$$R_4 \to R_4 + R_2$$

$$A = \begin{vmatrix} 5 & 9 & 6 & 0 \\ 0 & 4 & 1 & 0 \\ 0 & 4 & 1 & 0 \\ 0 & -4 & -1 & \frac{5}{12} \end{vmatrix}$$

$$R_3 \leftrightarrow R_4$$

$$A = \begin{bmatrix} 5 & 9 & 6 & 0 \\ 0 & 4 & 1 & 0 \\ 0 & 0 & 0 & \frac{5}{12} \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

$$R_2 \to \frac{R_2}{4}$$

$$R_3 \to R_3 \times \frac{12}{5}$$

$$A = \begin{bmatrix} 5 & 9 & 6 & 0 \\ 0 & 1 & \frac{1}{4} & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

$$R_1 \to R_1 - 9R_2$$

$$A = \begin{bmatrix} 5 & 0 & \frac{15}{4} & 0 \\ 0 & 1 & \frac{1}{4} & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

$$R_1 \to \frac{R_1}{5}$$

$$A = \begin{bmatrix} 1 & 0 & 3/4 & 0 \\ 0 & 1 & 1/4 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

$\therefore$ The matrix is not row equivalent to I_4.

$\therefore$ The polynomial does not form basis for P_3.

31. $H = Span\{V_1, V_2, V_3\}$ & $B = (V_1\ V_2\ V_3)$ Show that B is a basis for H & X is in H, find the B-coordinate vector of X fee.

$$V_1 = \begin{bmatrix} -6 \\ 4 \\ -9 \\ 4 \end{bmatrix} V_2 = \begin{bmatrix} 8 \\ -3 \\ 7 \\ -3 \end{bmatrix} V_3 = \begin{bmatrix} -9 \\ 5 \\ -8 \\ 3 \end{bmatrix} X = \begin{bmatrix} 4 \\ 7 \\ -8 \\ 3 \end{bmatrix}$$

Sol : **(Imp.)**

Given $H = Span\ (V_1\ V_2\ V_3)$

Consider the argumented matrix.

$$\begin{bmatrix} -6 & 8 & -9 & 4 \\ 4 & -3 & 5 & 7 \\ -9 & 7 & -8 & -3 \\ 4 & -3 & 3 & 3 \end{bmatrix}$$

$$R_2 \to 6R_2 + 9R_1$$

$$R_3 \to 6R_3 - 9R_1$$

$$R_4 \to 6R_4 + 4R_1$$

$$\begin{bmatrix} -6 & 8 & -9 & 4 \\ 0 & 14 & -6 & -38 \\ 0 & -30 & -33 & -84 \\ 0 & 14 & -18 & 34 \end{bmatrix}$$

$R_2 \to \frac{R_2}{2}$

$R_3 + \frac{R_3}{3}$

$R_4 \to \frac{R_4}{4}$

$$\begin{bmatrix} -6 & 8 & -9 & 4 \\ 0 & 7 & -3 & 29 \\ 0 & -10 & 11 & -28 \\ 0 & 7 & -9 & 17 \end{bmatrix}$$

$R_3 \to 7R_3 + 10R_2$

$R_4 \to R_4 - R_2$

$$= \begin{bmatrix} -6 & 8 & -9 & 4 \\ 0 & 7 & -3 & 29 \\ 0 & 0 & 47 & 94 \\ 0 & 0 & -6 & -12 \end{bmatrix}$$

$R_3 \to \frac{R_3}{47}$

$R_4 \to \frac{R_4}{6}$

$$\begin{bmatrix} -6 & 8 & -9 & 4 \\ 0 & 7 & -3 & 29 \\ 0 & 0 & 1 & 2 \\ 0 & 0 & 1 & 2 \end{bmatrix}$$

$R_1 \to R_1 + 9R_2$

$R_2 \to R_2 + 3R_3$

$R_4 + R_4 - R_3$

$$\begin{bmatrix} -6 & 8 & 0 & 22 \\ 0 & 7 & 0 & 35 \\ 0 & 0 & 1 & 2 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

$R_1 \to \frac{R_1}{-2}$

$R_2 \to \frac{R_2}{7}$

$$\begin{bmatrix} 3 & -4 & 0 & -11 \\ 0 & 1 & 0 & 5 \\ 0 & 0 & 1 & 2 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

$R_1 \to R_2 + 4R_2$

$$\begin{bmatrix} 3 & 0 & 0 & 9 \\ 0 & 1 & 0 & 5 \\ 0 & 0 & 1 & 2 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

$R_1 \to \frac{R_1}{3}$

$$\begin{bmatrix} 1 & 0 & 0 & 3 \\ 0 & 1 & 0 & 5 \\ 0 & 0 & 1 & 2 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

$\therefore$ The matrix contain pivot in each column

$\therefore$ B has a basis for H

Also by above system has a solution.

$\therefore$ B is a basis for H.

$\therefore$ The coordinate vector of B is

$$[X_B] = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ 3 \end{bmatrix} = \begin{bmatrix} 3 \\ 5 \\ 2 \end{bmatrix}$$

32. Let B = $\left\{\begin{bmatrix} 1 \\ -4 \end{bmatrix}\begin{bmatrix} -2 \\ 9 \end{bmatrix}\right\}$ Since the coordinate mapping determined by B is a linear trans formation from R^2 into R^2 this mapping must be implanted by some 2×2 matrix A find A.

Sol :

Given basis is $B = \left\{\begin{bmatrix}1\\-4\end{bmatrix}\begin{bmatrix}-2\\9\end{bmatrix}\right\}$

Hence $b_1 = \begin{bmatrix}1\\-4\end{bmatrix} b_2 = \begin{bmatrix}-2\\9\end{bmatrix}$

The matrix of transformation is

$PB^{-1} = [b_1 b_2]^{-1}$

$PB^{-1} = \begin{bmatrix}1 & -2\\-4 & 9\end{bmatrix}^{-1}$

The inverse of a matrix is given by

$A^{-1} = \frac{1}{ad-bc}\begin{bmatrix}d & -b\\-c & a\end{bmatrix}$

$PB^{-1} = \frac{1}{9-8}\begin{bmatrix}9 & 2\\4 & 1\end{bmatrix}$

$PB^{-1} = \begin{bmatrix}9 & 2\\4 & 1\end{bmatrix}$

$\therefore PB^{-1} = \begin{bmatrix}9 & 2\\4 & 1\end{bmatrix}$

33. Find the coordinates of a w.r.t basis set s = {x, y, z} where α **= (4, 5, 6).**

$$\mathbf{x} = \begin{bmatrix}\mathbf{1}\\\mathbf{1}\\\mathbf{1}\end{bmatrix} \quad \mathbf{y} = \begin{bmatrix}\mathbf{-1}\\\mathbf{1}\\\mathbf{1}\end{bmatrix} \quad \mathbf{z} = \begin{bmatrix}\mathbf{1}\\\mathbf{0}\\\mathbf{-1}\end{bmatrix}$$

why x, y, z $\in \mathbf{R^3}$

Sol :

Here the set,

S = {(1, 1, 1) (–1, 1, 1) (1, 0, –1)]

Since S is basis set so S is L.I and form a basis set.

Let $P \in R^3 \Rightarrow P = (a, b, c)$ where a, b, c $\in R$

(a, b, c) = *l*(x + my + nz)

(a, b, c) = l(1, 1, 1) + m(–1, 1, 1) + n(1, 0, –1) ... (1)

(a, b, c) = (*l*, *l*, *l*) + (–m, n, m) + (n, o, –n)

(a, b, c) = (*l* – m + n, *l* + m, *l* + m – n)

Equation on the both sides then we get

l – m + n = a ... (1)

l + n = b ... (2)

l + m – n = c ... (3)

Now putting there equations in matrix form

$$\left[\begin{array}{ccc|c}1 & -1 & 1 & a\\1 & 1 & 0 & b\\1 & 1 & -1 & c\end{array}\right]$$

$R_2 \to R_2 - R_1$

$R_3 \to R_3 - R1$

$$\left[\begin{array}{ccc|c}1 & -1 & 1 & a\\0 & 2 & -1 & b-a\\0 & 2 & -2 & c-a\end{array}\right]$$

$R_3 \to R_3 - R_1$

$$\left[\begin{array}{ccc|c}1 & -1 & 1 & a\\0 & 2 & -1 & b-a\\0 & 0 & -1 & c-a\end{array}\right]$$

Since the matrix is converted into echolon form then we get equations.

l – m + n = a

2m – n = b – a

– n = c – a

n = a – c

2m = b – a + n

2m = b – a + a – c

2m = b – c

$m = \frac{b-c}{2}$

2*l* = b + c

$$l = \frac{b-c}{2}$$

$$l - \left(\frac{b-c}{2}\right) + n = a$$

$2l - b + c + 2n = 2a$

$2l - b + c + 2a - 2c = 2a$

$2l - b - c = 0$

Now substituting the l, m, n values then we get,

$$(a, b, c) = \frac{b-c}{2}(1, 1, 1) + \frac{b-c}{2}(-1, 1, 1) + (a - c)(1, 0, 0)$$

Now the coordinates by consider,

$\alpha = (4, 5, 6)$

$$(4, 5, 6) = \frac{5+6}{2}(1, 1, 1) + \frac{5-6}{2}(-1, 1, 1) + (4 - 6)(1, 0, -1)$$

$$(4, 5, 6) = \frac{11}{2}(1, 1, 1) - \frac{1}{2}(-1, 1, 1) - 2(1, 0, -1)$$

$\therefore$ Coordinates are $\left(\frac{11}{2}, -\frac{1}{2}, -2\right)$ w.r.t to the bases set.

34. α = (1, 0, –1) w.r.t basis S ={x, y, z} where

$$\mathbf{x} = \begin{bmatrix} \mathbf{0} \\ \mathbf{1} \\ \mathbf{-1} \end{bmatrix} \quad \mathbf{y} = \begin{bmatrix} \mathbf{1} \\ \mathbf{1} \\ \mathbf{0} \end{bmatrix} \quad \mathbf{z} = \begin{bmatrix} \mathbf{1} \\ \mathbf{0} \\ \mathbf{2} \end{bmatrix}$$

Sol :

Here the set S = {(0, 1, –1) (1, 1, 0) (1, 0, 2)} is a basis set. So S is L.I and forms a spanning set.

Let $z = (a, b, c) \in R^3$

$(a, b, c) = p(0, 1, -1) + q(1, 1, 0) + r(1, 0, 2)$

$(a, b, c) = (q + r, p + q, -p + 2r)$

Now equating on both sides then we get

$q + r = a$

$p + q = b$

$-p + 2r = c$

Now taking into matrix form

$$\left[\begin{array}{ccc|c} 0 & 1 & 1 & a \\ 1 & 1 & 0 & b \\ -1 & 0 & 2 & c \end{array}\right]$$

$R_1 \leftrightarrow R_2$

$$\left[\begin{array}{ccc|c} 1 & 1 & 0 & b \\ 0 & 1 & 1 & a \\ -1 & 0 & 2 & c \end{array}\right]$$

$R_3 + R_3 + R_1$

$$\left[\begin{array}{ccc|c} 1 & 1 & 0 & b \\ 0 & 1 & 1 & a \\ 0 & 1 & 2 & c+b \end{array}\right]$$

$R_3 + R_3 - R_2$

$$\left[\begin{array}{ccc|c} 1 & 1 & 0 & b \\ 0 & 1 & 1 & a \\ 0 & 0 & 1 & c+b \end{array}\right]$$

Which has been converted into echolon form.

$p + q = b$

$q + r = a$

$r = c + b$

so $q = a - r$

$q = a - c - b$

$q = a - b - c$

$p = b - q$

$p = b - a + b + c$

$p = -a + 2b + c$

So

$(a, b, c) = -a + 2b + c\,(0, 1, -1) + (a - b - c)(1, 1, 0) + c + b\,(1, 0, 2)$

Now α = (1, 0, – 1) So

(1, 0, –1) = –1 + 0 – 1 (0, 1, –1)

+ (1, 0, + 1) (1, 1, 0)

+ –1 + 0 (1, 0, 2)

(1, 0, –1) = –2 (0, 1, –1) + 2(1, 1, 0)

– 1(1, 0, 2).

$\therefore$ The coordinates are (–2, 2, –1) w.r.t to the set S.

35. Find the coordinates of (6i, 7, 8i) w.r.t the basis s = {(1, 0, 0) (1, 1, 0) (1, 1, 1)} of $c^3(c)$.

Sol : **(Imp.)**

Given the set s = {(1, 0, 0) (1, 1, 0) (1, 1, 1)} is a basis set of $c^3(c)$.

Since the set 'S' is basis set SOS forms a L.I set.

Let z = (a, b, c) $\in C^3$ where a, b, c $\in$ C

Now (a, b, c) = p(1, 0, 0) + q(1, 1, 0) + r(1, 1, 1) where p, q, r $\in$ C.

$\Rightarrow$ (a, b, c) = (p + q + r, q + r, r)

$\Rightarrow$ (a, b, c) = (p + q + r, q + r, r)

Now equating on both sides we get,

p + q + r = a $\rightarrow$ (1)

q + r = b $\rightarrow$ (2)

r = c $\rightarrow$ (3)

Since equation (3) r = c so

q = b – r

q = b – c

p = q – q – r

So (a, b, c) = a – b (1, 0, 0) + b – c (1,1, 0) + c(1, 1, 1)

(a, b, c) = (a – b)(1, 0, 0) + (b – c)(1, 1, 0)

+ c(1, 1, 1)

z = L.C of the given vectors of S

Z $\in$ L(S)

$\therefore$ S is a basis of $c^3(c)$

So Now (a, b, c) = (6i, 7, 8i) then p = 6i – 7

q = 7 – 8i g = 8i which are the coordinates w.r.t the basis set 's'.

36. Let $V_1 = \begin{bmatrix} 1 \\ -2 \\ 3 \end{bmatrix}$ $V_2 = \begin{bmatrix} -2 \\ 7 \\ -9 \end{bmatrix}$ Now determine if $\{V_1, V_2\}$ is a basis for R^3. Is $\{V_1, V_2\}$ is a basis for R^2?

Sol :

The given vector are $V_1 = \begin{bmatrix} 1 \\ -2 \\ 3 \end{bmatrix}$ $V_2 = \begin{bmatrix} -2 \\ 7 \\ 9 \end{bmatrix}$. Now to verify the $S = \{V_1, V_2\}$ is abasis we have to show that S is L.I and linear spam of V.

Let us consider the scalars $a,b \in R$ and

$\alpha_1 = (1, -2, 3)$ and $\alpha_2 = (-2, 7, -9)$

where α_1, α_2 and R^3.

Such that $a\alpha_1 + b\alpha_2 = \bar{0}$

$\Rightarrow$ $a(1, -2, 3) + b(-2, 7, -9) = (0, 0, 0)$

$\Rightarrow$ $(a, -2a, 3a) + (-2b, -7b, -9b) = (0, 0, 0)$

$\Rightarrow$ $(a - 2b, -2a - b, 3a - 9b) = (0, 0, 0)$

$a - 2b = 0$ .... (1)

$-2a - 7b = 0$ .... (2)

$3a - ab = 0$ .... (3)

$$\begin{bmatrix} 1 & -2 \\ -2 & -7 \\ 3 & -9 \end{bmatrix}$$

Now converting the above matrix in the echolon form then we get.

$$\begin{bmatrix} 1 & -2 \\ -2 & -7 \\ 3 & -9 \end{bmatrix}$$

$R_2 \to R_2 + 2R_1$

$R_3 \to R_3 + 2R_1$

$$\begin{bmatrix} 1 & -2 \\ 0 & -11 \\ 0 & -3 \end{bmatrix}$$

$R_3 \to 11R_3 - 3R_2$

$$\begin{bmatrix} 1 & -2 \\ 0 & -11 \\ 0 & -3 \end{bmatrix}$$

Since it is converted into echolon form so consider the equations.

$a - 2b = 0$

$-11b = 0$

$\Rightarrow \quad b = 0$

$a = 2b = 0$

$\therefore a = 0, b = 0$

Since the scalars are equal to zero so the set 'S' is L.I set.

Let $z = (a, b, c) \in R^3$ such that

$(a, b, c) = p(1, -2, 3) + q(-2, 7, 9)$

$\Rightarrow \quad (a, b, c) = (p - 2q, -2p + 7q, 3p + 9q)$

Now equating on both sides we get

$p - 2q = a \quad ... (1)$

$-2p + 7q = b \quad ... (2)$

$3p + 9q = c \quad ... (3)$

Now putting in the matrix form

$$\begin{bmatrix} 1 & -2 \\ -2 & 7 \\ 3 & 9 \end{bmatrix}$$

$R_2 \to R_2 + 2R_1$

$R_3 \to R_3 + 3R_1$

$$\begin{bmatrix} 1 & -2 \\ 0 & 3 \\ 0 & 15 \end{bmatrix}$$

$R_3 \to R_3 - 2R_2$

$$\begin{bmatrix} 1 & -2 \\ 0 & 3 \\ 0 & 0 \end{bmatrix}$$

Now the above matrix has been converted into echolon form so taking the equations.

$p - 2q = 0$

$3q = 0$

$\therefore p = 0$ and $q = 0$

So $\{V_1 V_2\}$ does not form basis of R^2.

UNIT II

Probability - Basic terminology, Three types of probability, Probability rules, Statistical independence, statistical dependency, Bayes' theorem.

Probability Distributions - Random variables, expected values, binomial distribution, Poisson distribution, normal distribution, choosing correct distribution.

2.1 PROBABILITY

Q1. Define the term probability.

Ans :

The theory of probability is one of the most useful and interesting branches of modem mathematics. It is becoming prominent by its application in many fields of learning, such as Insurance, Statistics, Biological Sciences, Physical Sciences, Engineering, etc.

An Italian mathematician, Galileo (1564 - 1642), attempted a quantitative measure of probability while dealing with some problems related to gambling. In the middle of 17th Century, two French mathematicians, Pascal and Fermat, laid down the first foundation of the mathematical theory of probability while solving the famous 'Problem of Points' posed by Chevalier-De- Mere. Other mathematicians from several countries also contributed in no small measure to the theory of probability. Outstanding of them were two Russian mathematicians, A. Kintchine and A. Kolmogoroff, who axiomised the calculus of probability.

If an experiment is repeated under similar and homogeneous conditions, we generally come across two types of situations.

(i) The net result, what is generally known as 'outcome' is unique or certain.

(ii) The net result is not unique but may be one of the several possible outcomes.

The situations covered by

(i) are known as 'deterministic' or 'predectable' and situations covered by

(ii) are known as 'probabilistic' or 'un-predictable'.

'Deterministic' means the result can be predicted with certainty.

For example, if r is the radius of the sphere then its volume is given by $V = \frac{4}{3}\pi r^3$ which gives uniquely the volume of the sphere.

There are some situations which do not lend themselves to the deterministic approach and they are known as 'Probabilistic'.

For example, by looking at the sky, one is not sure whether the rain comes or not.

In such cases we talk of chances or probability which can be taken as a quantitative measure of certainty. We will now introduce the concept of probability with definitions.

2.1.1 Basic terminology

Q2. Explain the basic terminology are used in probability.

Ans : **(Imp.)**

1. Random Experiment

Random experiment refers to a situation where erratic, nonspecific and arbitrary results can be obtained.

If an experiment is conducted repeatedly for V number of times under identical conditions and if the outcome of an experiment varies in each case, then such an experiment is called as 'random experiment'.

Example: Tossing a coin, rolling a dice etc.

2. Outcome

The result of a random experiment is usually referred as an outcome.

If we toss a coin, the outcome may be a head or a tail. In such a case, number of outcomes = 2.

3. Event

An event is a possible outcome of an experiment or a result of trial.

Basically there are two types of events. Simple and compound event.

(a) Simple Event: The probability of happening or non-happening of a single event is considered as a simple event.

Example: When we are selecting two black coins from a box containing 10 white and 5 black coins.

(b) Compound Event: If the joint occurrence of two or more event is considered then it is known as 'compound event' or 'composite event'.

Example: A box containing 5 white balls, 3 black balls and 8 red balls, when we draw 2 white balls in first draw, 2 black balls in second draw and 5 red balls in third draw.

4. Mutually Exclusive Events

When two events cannot occur simultaneously in a single trial then such events are called as mutually exclusive or incompatible events.

Example

(a) When a single coin is tossed, either a head or a tail can turn up, both cannot come at the same time

(b) At a single point of time, a person can be alive or dead.

5. Non-mutually Exclusive Events

When two events can occur simultaneously in a single trial then such events are said to be non-mutually exclusive events.

Example: From a pack of cards, drawing a red card and drawing a queen are the two events. These two events can occur simultaneously while drawing a red queen.

Hence, these two events are said to be non-mutually exclusive events which can occur at the same time.

6. Collectively Exhaustive Events

Collectively exhaustive events are those events whose totality contains all the potential outcomes of a random experiment.

Example: When a dice is thrown, the total possible outcomes are 1, 2, 3, 4, 5 and 6 and thus the number of exhaustive cases is 6.

7. Equally Likely Events

Events are considered as equally likely events when the probability of occurrence of all the events is equal.

These events are also called as 'equally, probable events'.

Example: If a coin is tossed, the two possible outcomes are head and tail. The probability of their occurrence is equal (i.e,

8. Independent Events

Two or more events are considered as independent events, when the outcome of one event does not influences and is not influenced by the other event.

Example: When a student has appeared in physics and chemistry examinations, his marks obtained in physics is independent of the marks obtained in chemistry.

9. Dependent Events

Dependent events are the events in which the occurrence or non-occurrence of one event in any one trial influences the probability of occurrence of other events in other trials.

Example: When a card is drawn from a pack of playing cards and is not replaced then this changes the probability of occurrence of the second card.

Here probability of occurrence of second event is dependent on the occurrence of first event.

10. Complementary Events

Two events are said to be complementary events if they are mutually exclusive and collectively exhaustive.

Example: When a coin is tossed, getting a head or a tail are mutually exclusive and collectively exhaustive.

Hence, if we get tail then head is considered as its complementary event.

2.1.2 Three Types of Probability

Q3. State the three types of probabilities with an examples.

Ans :

1. Theoretical Probability
2. Experimental Probability
3. Axiomatic Probability

1. Theoretical Probability

It is based on the possible chances of something to happen.

Theoretical probability is the theory behind probability. To find the probability of an event using theoretical probability, it is not required to conduct an experiment. Instead of that, we should know about the situation to find the probability of an event occurring. The theoretical probability is defined as the ratio of the number of favourable outcomes to. the number of possible outcomes.

For example, if a coin is tossed, the theoretical probability of getting head will be ½

Probability of Event P(E) = No. of. Favourable outcomes/ No. of. Possible outcomes.

Example

Find the probability of rolling a 5 on a fair die.

Sol :

To find the probability of getting 5 while rolling a die, an experiment is not needed. We know that, there are 6 possible outcomes when rolling a die. They are 1, 2, 3, 4, 5, 6.

Therefore, the probability is,

Probability of Event P(E) = No. of. Favourable outcomes/ No. of Possible outcomes.

P(E) = 1/6.

Hence, the probability of getting 5 while rolling a fair die is 1/6.

2. Experimental Probability

It is based on the basis of the observations of an experiment. The experimental probability can be calculated based on the number of possible outcomes by the total number of trials. For example, if a coin is tossed 10 times and heads is recorded 6 times then, the experimental probability for heads is 6/10 or, 3/5.

The experimental Probability is defined as the ratio of the number of times that event occurs to the total number of trials.

Probability of Event P(E) = No. of. times that event occurs/ Total number of trials

The basic difference between these two approaches is that in the experimental approach; the probability of an event is based on what has actually happened by conducting a series of actual experiments, while in theoretical approach; we attempt to predict what will occur without actually performing the experiments.

Example 1:

If a coin is tossed 10 times, head appears 3 times. Find experimental probability of getting a head.

Sol :

Experimental probability of getting a head = 3/10

Example 2:

A survey was conducted to determine students' favorite brands of sneakers.

Each student chose only one brand from the list of brands A, B, C, D, or E.

What is the probability that a student's favorite sneaker was brand D?

Sneaker	A	B	C	D	E
Number	12	15	24	26	13

Sol :

There were 12 + 15 + 24 + 26 + 13 = 90 "trials" in this experiment (each student's response was a trial).

26 out of the 90 students chose brand D.

The probability is :

$$\frac{\text{Number choosing brand D}}{\text{Total number chossing a brand}} = \frac{26}{90} = \frac{13}{45}$$

3. Axiomatic Probability

In axiomatic probability, a set of rules or axioms are set which applies to all types. These axioms are set by Kolmogorov and are known as Kolmogorov three axioms. With the axiomatic approach to probability, the chances of occurrence or non-occurrence of the events can be quantified. The axiomatic probability less on covers this concept in detail with Kolmogorov's three rules (axioms) along with various examples.

Conditional Probability is the likelihood of an event or outcome occurring based on the occurrence of a previous event or outcome.

Probability is a set function P(E) that assigns to every event E a number called the "Probability of E" such that:

1. The probability of an event is greater than or equal to zero.

$$P(E) \geq 0$$

2. The probability of the sample space is one

$$P(\Omega) = 1$$

One important thing about probability is that it can only be applied to experiments where we know the total number of outcomes of the experiment, i.e. unless and until we know the total number of outcomes of an experiment, concept of probability cannot be applied.

Thus, in order to apply probability in day to day situations, we should know the total number of possible outcomes of the experiment. Axiomatic Probability is just another way of describing the probability of an event. As, the word itself says, in this approach, some axioms are predefined before assigning probabilities. This is done to quantize the event and hence to ease the calculation of occurrence or non-occurrence of the event.

Axiomatic Probability Conditions

Let, S be the sample space of any random experiment and let P be the probability of occurrence of any event. Noting the characteristics of P, it should be a real valued function whose domain will be the power set of S and the range will lie in the interval [0,1]. This probability P will satisfy the following probability axioms:

1. For any event E, $P(E) \geq 0$
2. P(S) = 1
3. In case E and F are mutually exclusive events, then following equation will be valid: $P(E \cup F) = P(E) + P(F)$

From point (3) it can be stated that $P(\phi) = 0$

If we need to prove this, let us take $F = \phi$ and make a note that

E and ϕ are disjoint events. Hence, from point (3) we can deduce that-

$P(E \cup \phi) = P(E) + P(\phi)$ or

$P(E) = P(E) + P(\phi)$

i.e. $P(\phi) = 0$

Let, the sample space of S contain the given outcomes $\delta_1, \delta_2, \delta_3 .... \delta_n$, then as per axiomatic definition of probability, we can deduce the following points-

1. $0 \leq P(\delta_i) \leq 1$ for each $\delta_i \in S$
2. $P(\delta_1) + P(\delta_2) + ... + P(\delta_n) = 1$
3. For any event Q, $P(Q) = \Sigma P(\delta_i), \delta_i \in Q$

This is to be noted that the singleton $\{\delta_i\}$ is known as elementary event and for the convenience of writing, we write $P(\delta_i)$ for $P(\{\delta_i\})$.

Example

On tossing a coin we say that the probability of occurrence of head and tail is 12 each. Basically here we are assigning the probability value of 12 for the occurrence of each event.

Sol :

This condition basically satisfies both the conditions, i.e.

- Each value is neither less than zero nor greater than 1 and
- Sum of the probabilities of occurrence of head and tail is 1

Hence, for this case we can say that the probabilities of occurrence of head and tail are 12 each.

Now, say P(H) = 58 and P(T) = 38

Does this probability value satisfy the conditions of axiomatic approach?

For this, let us again check the basic initial conditions of the axiomatic approach of probability.

- Each value is neither less than zero nor greater than 1 and
- Sum of the probabilities of occurrence of head and tail is 1

Hence this sort of probability value assignment also satisfies the axiomatic approach of probability. Thus, we can conclude that there can be infinite ways to assign the probability to outcomes of an experiment.

PROBLEMS

1. What is the probability that a card drawn at random from the pack of paying cards may be either a queen or a king.

Sol :

Let S be the sample space associated with the drawing of a card.

$\therefore \quad n(S) = {}^{52}C_1 = 52$

Let E_1 be the event of the card drawn being a queen.

$\therefore \quad n(E_1) = {}^{4}C_1 = 4$

Let E_2 be the event of the card drawn being a king.

$\therefore \quad n(E_2) = {}^{4}C_1 = 4$

But E_1, E_2 are mutually exclusive events.

Since $E_1 \cup E_2$ is the event of drawing either a queen or a king, we have

$$P(E_1 \cup E_2) = P(E_1) + P(E_2) = \frac{n(E_1)}{n(S)} + \frac{n(E_2)}{n(S)} = \frac{4}{52} + \frac{4}{52} = \frac{2}{13}$$

2. In a group there are 3 men and 2 women. Three persons are selected at random from this group. Find the probability that one man and two women or two men and one women are selected.

Sol :

Let S be the sample space associated with the selection of 3 persons out of 5.

$\therefore \quad n(S) = {}^{5}C_5 = 10$

Let E_1 be the event of selecting 1 men and 2 women

$\therefore \quad n(E_1) = {}^{3}C_1 \times {}^{2}C_2$

Let E_2 be the event of selecting 2 men and a women

$\therefore \quad n(E_2) = {}^{3}C_2 \times {}^{2}C_1 = 6$

But $E_1 \cap E_2 = \phi$, i.e., E_1, E_2 are mutually exclusive

Now $E_1 \cup E_2$ is the event of selecting 1 man and 2 women or 2 men or 1 women.

$$\therefore \quad P(E_1 \cup E_2) = P(E_1) + P(E_2) \frac{n(E_1)}{n(S)} + \frac{n(E_2)}{n(S)} = \frac{3}{10} + \frac{6}{10} = \frac{9}{10}$$

3. In a sample of 446 cards, stopped at a road block, only 67 of the drivers, has their seat belts fastered. Estimate the probability that a driver stopped on that road, will have his or her seat belt fastered.

Sol :

Number of favourable outcomes = 67

Exhaustive events = 446

$\therefore$ Required probability = $\frac{67}{446}$

4. Twelve balls are distributed or random among three boxes. What is the probability that the first box will contain 3 balls ?

Sol :

Since each ball can go to any one of the three boxes there are three ways in which a ball can go to any one of the three boxes. Hence there are 3^{12} ways in which 12 balls can be placed in the three boxes.

Number of ways in which 3 balls out of 12 can go to the first box is $^{12}C_3$.

Now the remaining 9 balls are to be placed in remaining 2 boxes and their can be done in 2^9 ways.

$\therefore$ Total number of favourable cases = $^{12}C_3 \times 2^9$

$\therefore$ Required probability = $\frac{^{12}C_3 \times 2^9}{3^{12}} = 0.212$

2.1.3 Probability Rules

Q4. Explain addition theorem of probability.

Ans : **(Imp.)**

Addition theorem is different for mutually exclusive events.

(i) For Mutually Exclusive Events

When 'A' and 'B' are two mutually exclusive events (i.e., both cannot occur at the same time) then the probability of occurrence of A or B is equal to the sum of their individual probabilities.

$$\boxed{\begin{aligned} &P(A \text{ or } B) = P(A) + P(B) \\ \Rightarrow\ &P(A \cup B) = P(A) + P(B) \end{aligned}}$$

Diagrammatically it can be represented as,

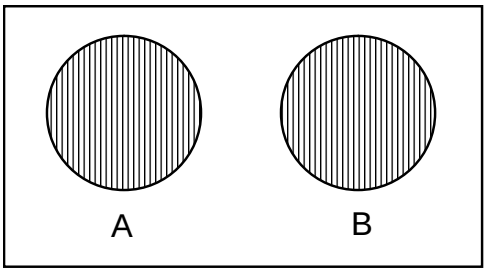


Fig.: Mutually Exclusive Events

In case of 3 events A, B and C,

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

(ii) For Non-mutually Exclusive Events

In case of non-mutually exclusive event (i.e., if the events occur together) these is a variation in the addition theorem.

When 'A' and 'B' are non-mutually exclusive events then the probability of occurrence of A or B is the sum of their individual probability which should be deducted from the probability of A and B occurring together.

$$\boxed{P(A \text{ or } B) \Rightarrow P(A \cup B) = P(A) + P(B) - P(P \cap B)}$$

Diagrammatically it can be represented as,

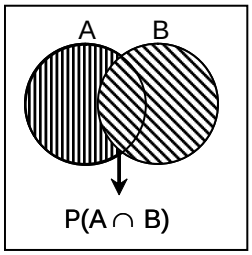


Fig.: Non-Mutually Exclusive Events

In case of three non-mutually exclusive events.

A, B and C the probability of occurrence of A or B or C and be calculated by the following formula,

$$\boxed{P(A \text{ or } B \text{ or } C) = P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)}$$

PROBLEMS

5. The probability that a contractor will get a plumbing contract is 3/4 and the probability that he will not get electric contract is 4/9. If the probability of getting at least one contract is 5/6, what is the probability that he will get both the contracts?

Sol :

Let 'A' be the event that the contractor will get plumbing contract.

'B' be the event that the contractor will get electric contract.

The probability of getting plumbing contract = P(A)

The probability of not getting a plumbing contract = $P(\bar{A})$

The probability of getting electric contract = P(B)

The probability of not getting electric contract = $P(\bar{B})$

The probability of getting at least one contract is,

$$P(A \text{ or } B) = \frac{5}{6}$$

Given that,

$$P(A) = \frac{3}{4}; P(\bar{A}) = \frac{1}{4} \quad (\because P(A) + P(\bar{A}) = 1)$$

$$P(\bar{B}) = \frac{4}{9}; P(B) = \frac{5}{9}$$

$\therefore$ The probability that he will get both the contracts can be given by using addition theorem.

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

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

$$= \frac{3}{4} + \frac{5}{9} - \frac{5}{6}$$

$$= \frac{27 + 20 - 30}{36} = \frac{47 - 30}{36} = \frac{17}{36} = 0.472 = 47.2\%$$

$\therefore$ The probability that the contractor will get both the contracts $= \frac{17}{36} = 0.472 = 47.2\%$.

6. A card is drawn from a standard pack of 52 cards.

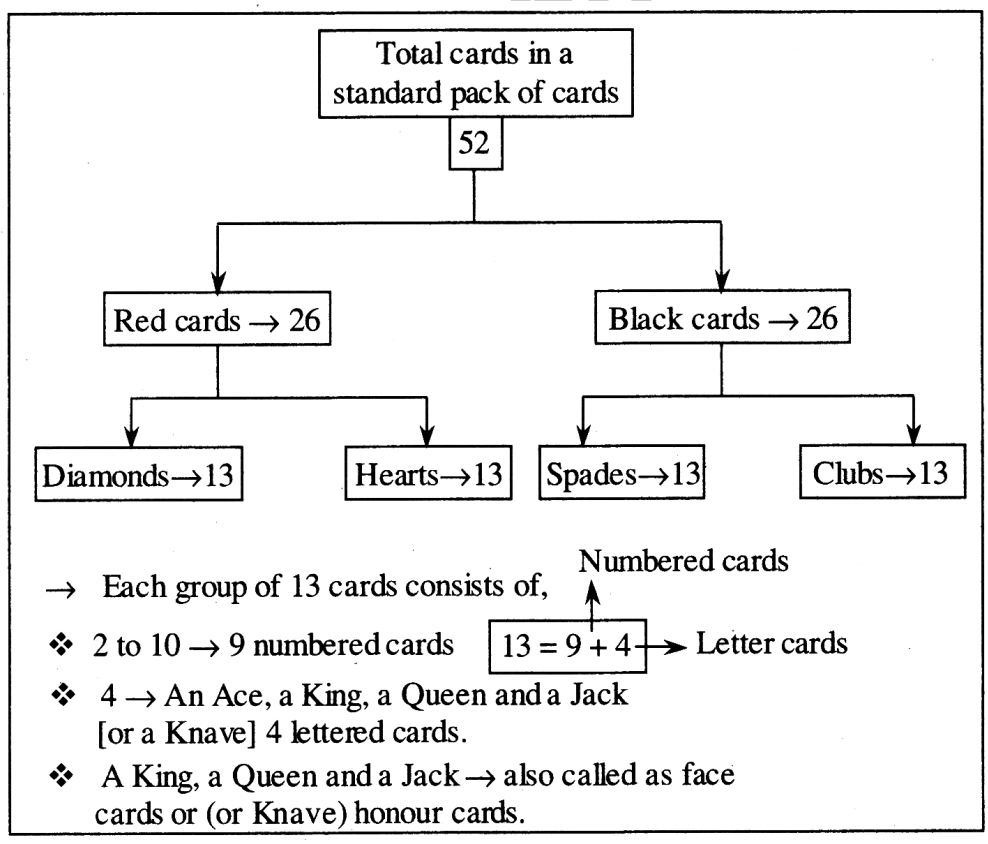
(i) What is the probability that it is a black or a red card?

(ii) What is the probability that it is an Ace or a King or a Queen?

(iii) What is the probability that it is a red Knave or a black King?

Sol :

Before solving the problem, one should know the entire picture of standard pack of cards.



Note

In a standard pack of 52 cards,

- There are 13 diamonds, 13 hearts, 13 spades and 13 clubs.
- There are 26 red cards and 26 black cards.
- In total, there are 36 numbered cards and 16 lettered cards.
- There are 12 face cards (honour cards) in a pack of 52 cards, i.e., 4 Kings, 4 Queens and 4 Jacks (Knaves).
- There are 4 Aces in total in a standard pack of cards.

(i) Probability of Getting a Black or a Red Card

Let,

'A' be the event of getting a black card

'B' be the event of getting a red card.

Probability of getting a black card,

$$P(A) = \frac{26}{52} = \frac{1}{2}$$

Probability of getting a red card,

$$P(B) = \frac{26}{52} = \frac{1}{2}$$

Since, the events "A" and "B" are mutually exclusive events, the probability of getting a black or red card.

$$\boxed{P(A \cup B) = P(A) + P(B)}$$

[By addition theorem of probability]

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

(ii) Probability of Getting an Ace or a King or a Queen

As we know, there are 4 Aces, 4 Kings and 4 Queens in a pack of 52 cards.

Let,

'A' be the event of getting an Ace

'B' be the event of getting a King

'C' be the event of getting a Queen.

Probability of getting an Ace,

$$P(A) = \frac{4}{52} = \frac{1}{13}$$

Probability of getting a King,

$$P(B) = \frac{4}{52} = \frac{1}{13}$$

Probability of getting a Queen,

$$P(C) = \frac{4}{52} = \frac{1}{13}$$

Since A, B, C are mutually exclusive events, the probability of getting an Ace or a King or a Queen is given by,

$$\boxed{P(A \cup B \cup C) = P(A) + P(B) + P(C)}$$

[By addition theorem of probability]

$$= \frac{1}{13} + \frac{1}{13} + \frac{1}{13} = \frac{3}{13}$$

$\therefore$ The probability of getting on Ace on a King or a Queen is $\frac{3}{13}$

(iii) Probability of Getting a Red Knave or a Black King

As we know, there are 26 red cards and 26 black cards in a pack of 52 cards.

Out of 26 red cards, there are 2 Knaves.

There are 2 Kings out of 26 black cards.

Let,

'A' be the probability of getting a red Knave and

'B' be the probability of getting a black king.

Probability of getting a red Knave,

$$P(A) = \frac{2}{26} = \frac{1}{13}$$

Probability of getting a black King,

$$P(B) = \frac{2}{26} = \frac{1}{13}$$

As 'A' and 'B' are mutually exclusive events, the probability of getting a red Knave or a black King is given by,

$$\boxed{P(A \cup B) = P(A) + P(B)}$$

[By addition theorem of probability for mutually exclusive events].

$$\therefore \quad P(A \cup B) = \frac{1}{13} + \frac{1}{13} = \frac{2}{13}$$

$\therefore$ The probability of getting a red Knav or a black king $= \frac{2}{13}$.

7. One card is drawn from a standard pack of 52. What is the probability that it is either King or a Queen?

Sol :

Let *'A'* be the event of drawing a king from a standard pack of 52.

'B' be the event of drawing a queen.

P(A) = The probability of drawing a king

P(B) = Probability of drawing a queen.

As, the probability of drawing either a king or a queen has to be determined by drawing a single card, the events are said to be mutually exclusive.

For mutually exclusive events,

$$\boxed{P(A \cup B) = P(A) + P(B)}$$

(By addition theorem of probability)

Given data,

$P(A) = \frac{4}{52}$ (As there are 4 kings in a standard pack of 52).

$P(B) = \frac{4}{52}$ (As there are 4 queens in a standard pack of 52).

$$\therefore \quad P(A \cup B) = \frac{4}{52} + \frac{4}{52} = \frac{8}{52} = \frac{2}{13}$$

Therefore, the probability of drawing either a king or a queen $= \frac{2}{13}$.

2.2 Probabilities under Conditions of Statistical Independence

Q5. Write about the various probabilities under statistical independence.

Ans : **(Imp.)**

Statistical independent events are those events, whose occurrence has no effect on the occurrence of any other event.

The happening or not happening of any event does not affect the happening of another event.

Example

If a coin is tossed twice, the result of second tossing would not be affected by the result of the first tossing.

There are three types of probabilities under statistical independence

1. Marginal probability
2. Joint probability
3. Conditional probability
   (a) For independent events
   (b) For dependent events.

(i) Marginal Probabilities under Statistical Independence

Marginal probability is the simple probability of the occurrence of an event.

It is also known as 'unconditional' probability or 'single probability'.

For example, when a single unbiased coin is tossed the probability that it is a head is P(H) $= \frac{1}{2}$ or probability that it is a tail, P(T) $= \frac{1}{2}$.

These two events are independent and do not overlap one another i.e., the events are statistically independent of the outcomes of next coin been tossed.

The individual probabilities obtained in this case i.e., P(H) or P(T) are called as marginal probabilities.

(ii) Joint Probabilities Under Statistical Independence

A joint probability is the probability of occurrence of two or more simple independent events together.

In other words, the product of two marginal probabilities occurring together or in succession is called as 'joint probability'.

For example, let P(W) is the probability of a girl with white complexion and P(B) is the probability of a girl with black hair.

Here, P(W) and P(B) are marginal probability.

Probability of a girl with white complexion and with black hair together is called as 'joint probability'.

Joint probability of these two events can be represented as, $P(W \cap B)$.

$$\therefore P(W \cap B) = P(W).P(B)$$

(By multiplication theorem)

$P(W \cap B) \rightarrow$ Where, Probability of a girl with white complexion and with black hair.

P(W) = Probability of a girl with white complexion

P(B) = Probability of a girl with black hair.

For example, there are 100 girls of whom 60 girls are white and 70 girls possess black hair, the probability of girls with white complexion and black hair is called joint probability.

Mathematically, it can be calculated as follows,

$$P(W) = \frac{60}{100} = 0.6$$

$$P(B) = \frac{70}{100} = 0.7$$

Probability of a girl having white complexion as well as black hair,

$$P(W \cap B) = 0.6 \times 0.7 = 0.42 \text{ or } 42\%$$

Here, joint probability = 0.42.

Similarly, joint probability for three independent events A, B and C can be represented as $P(A \cap B \cap C)$.

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

(By multiplication theorem of probability).

(iii) Conditional Probabilities under Statistical Independence

Conditional probability is the probability of occurrence of second event (B) given that the first event (A) has already occurred.

For statistically independent events, the conditional probability of event B given that event A has already occurred is simply the probability of event B. Symbolically, it can be represented as,

$$P(B/A) = P(B)$$

and is read as the probability of event *B,* given that the event A has already occurred.

For example, the probability of a second toss of a fair coin given that the tossing of first coin resulted as head.

In the given example, occurrence of first toss is the event A and occurrence of second toss is the event B.

The outcome of first toss is head which has absolutely no effect on the result of the second toss.

In this case, the conditional probability of event *B* given that A has already occurred is equal to the probability of event *B.*

Note

Conditional Probability for Statistically Dependent Events.

If the occurrence of one event affects the occurrence of another event then the two events are said to be "statistically dependent events".

When the probability of one event is dependent or affected by the occurrence of some other event, then it is called as 'conditional probability for statistically dependent events'.

Let the event A has already occurred and A and B are two dependent events, then the probability of occurrence of B given A has already occurred is represented as P(BIA).

$$\therefore P(B / A) = \frac{P(A \cap B)}{P(A)}$$

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

= Probability of occurrence of both the events A and B.

P(A) = Marginal probability

= Probability of occurrence of A

Note

Conditional probability for statistically dependent events is joint probability divided by marginal probability in this case,

(OR)

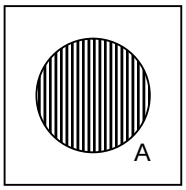
Joint probability is the product of conditional probability and marginal or prior probability.

Example

A person has appeared to an interview. The person posses an MBA degree the probability that the person getting selected in the interview given that the person is a MBA holder can be regarded as the example for conditional probability for dependent events.

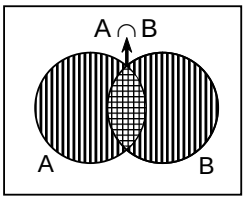
1. Marginal Probability

P(A)



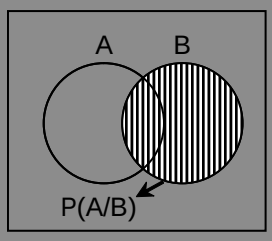
2. Joint Probability

$P(A \cap B)$



3. Conditional Probability

(a) For dependent events P(A/B).



(b) For independent events P(A/B) = P(A)

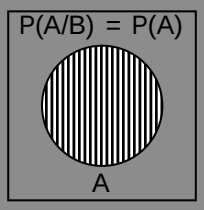


Fig.: Diagrammatic Representation of Various Probabilities

PROBLEMS

8. A restaurant features two types of lunches special lunch and ordinary lunch. Sixty percent of its men customers order the special lunch and the rest of the men order the ordinary lunch. Ninety percent of its women customers order the ordinary lunch and the rest order the special lunch. Seventy percent of its customers are men. In its preparations, what ratios of special to ordinary lunches should the restaurant plan for? Determine the probability that the man will order a special lunch.

Sol :

Let 'M' denote the event of the customer being a man.

'W indicate the event of the customer being a women.

'S' indicate the event of ordering a special lunch.

'O' indicate the event of ordering ordinary lunch.

Given that,

The probability of men who ordered special lunch,

$P(M \cap S) = 60\% = 0.6$

Probability of men ordering ordinary lunch,

$P(M \cap O) = 40\% = 0.4.$

Probability of women ordering ordinary lunch,

$P(W \cap O) = 90\% = 0.9$

Probability of women ordering special lunch,

$P(W \cap S) = 10\% = 0.1$

Probability of customers who are men,

$P(M) = 70\% = 0.7$

Probability of customers who are women,

$P(W) = 1 - P(M) = 1 - 0.7 = 0.3.$

Let us consider total 100 customers of whom 70 are men and 30 are women. The ratio of men to women = 70 : 30.

With the help of the given information the percentages can be converted to numbers.

Total customers = 100

Total men = 70

Total women = 30.

Men ordering special lunch = 60% of 70 = $\frac{60}{100} \times 70 = 42$

Men ordering ordinary luch = 40% of 70 = $\frac{40}{100} \times 70 = 28$

Women ordering special lunch = 10% of 30 = $\frac{10}{100} \times 30 = 3$

Women ordering ordinary luch = 90% of 30 = $\frac{90}{100} \times 30 = 27$

Tabulating the above results.

	S	O	Total
M	42	28	70
W	3	27	30
Total	45	55	100

(i) Therefore, the ratio of special lunch to ordinary lunch from the above table,

$$S : O = 45 : 55$$
$$= 9 : 11$$

(ii) Probability that the man will order a special launch = P(S/M).

$$P(S/M) = \frac{P(S \cap M)}{P(M)} \text{ (Conditional Probability)}$$

$\therefore$ $P(S \cap M)$ = Probability of a man and a special lunch

$$= \frac{42}{100} = 0.42 \text{ (From table)}$$

$P(M)$ = Probability of a customer being a man

$$= \frac{70}{100} = 0.7 \text{ (From table)}$$

$$\therefore P(S/M) = \frac{0.42}{0.70} = 0.6$$

Therefore, the probability that a man will order a special lunch = 0.6 or 60%.

2.3 Probabilities under Conditions of Statistical Dependency

Q6. Explain briefly about Probabilities under Conditions of Statistical dependency.

Ans :

Statistical dependence exists when the probability of some event is dependent on or affected by the occurrence of some other event. Just as with independent events, the types of probabilities under statistical dependence are

1. Conditional
2. Joint
3. Marginal

1. Conditional Probabilities under Statistical Dependence

Conditional and joint probabilities under statistical dependence are more involved than marginal probabilities are

Assume that we have one box containing 10 balls distributed as follows :

- Three are colored and dotted
- One is colored and striped

- Two are gray and dotted
- Four are gray and striped

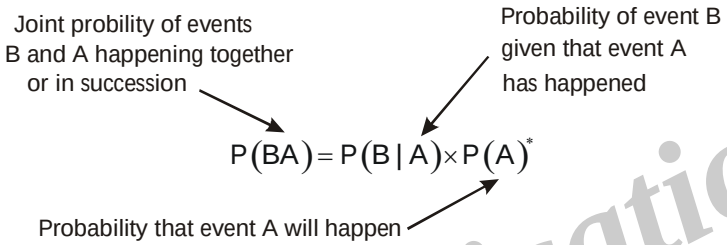
The probability of drawing any one ball from this box is 0.1, since there are 10 balls, each with equal probability of being drawn.

Conditional Probability for Statistically Dependent Events

$$P(B|A) = \frac{P(BA)}{P(A)}$$

2. Joint Probabilities under Statistical Dependence

Joint Probability for Statistically Dependent Events



Notice that this formula is not P(BA) = P(B) × P(A), as it would be under conditions of statistical independence.

3. Marginal Probabilities under Statistical Dependence

Marginal probabilities under statistical dependence are computed by summing up the probability abilities of all the joint events in which the simple event occurs.

Q7. State the Probabilities under Statistical Independence and Dependence

Ans :

Type of Probability	Symbol	Formula under Statistical Independence	Formula under Statistical Dependence
Marginal	P(A)	P(A)	Sum of the probabilities of the Joint events in which A occurs
Joint	P(AB) or P(BA)	P(A) × P(B) P(B) × P(A)	P(A|B) × P(B) P(B|A) × P(A)
Conditional	P(B|A)	P(B)	$\frac{P(BA)}{P(A)}$
	or P(A|B)	P(A)	$\frac{P(AB)}{P(B)}$

2.4 BAYES' THEOREM

Q8. State the explain Bayes' theorem.

Ans : **(Imp.)**

$E_1, E_2, ..., E_n$ are n mutually exclusive and exhaustive events such that $P(E_1) > 0$ (i = 1, 2, ..., n) in a sample space S and A is any other event in S intersecting with every E_i (i.e., A can only occur in combination with any one of the events $E_1, E_2, ..., E_n$) such that P(A) > 0.

If E_i is any of the events of $E_1, E_2, ..., E_n$ where $P(E_1), P(E_2), ..., P(E_n)$ and $P(A/E_1), P(A/E_2), ..., P(A/E_n)$, are know, then

$$P(E_K / A) = \frac{P(E_K) \cdot P(A / E_K)}{P(E_1) \cdot P(A / E_1) + P(E_2) \cdot P(A / E_2) + ... + P(E_n) \cdot P(A / E_n)}$$

Proof :

$E_1, E_2, ..., E_n$ are n events of S such that $P(E_i) > 0$ and $E_i \cap E_j = \phi$ for $i \neq j$ where i, j = 1, 2, ..., n. Also $E_1, E_2, ..., E_n$ are exhaustive events of S and A is any other event of S where P(A) > 0.

$$S = E_1 \cup E_2 \cup ... \cup E_n \text{ and}$$

$$A = A \cap S = A \cap (E_1 \cup E_2 \cup ..... \cup E_n)$$

$$= (A \cap E_1) \cup (A \cap E_2) \cup ..... \cup (A \cap E_n)$$

Here $A \cap E_1, A \cap E_2, ..., A \cap E_n$, are mutually exclusive events. Then

$$P(E_K / A) = \frac{P(E_K - \cap A)}{P(A)} = \frac{P(E_K \cap A)}{P[(A \cap E_1) \cup P(A \cap E_2) \cup ..... \cup P(A \cap E_n)]}$$

$$= \frac{P(E_K \cap A)}{P(A \cap E_1) + P(A \cap E_2) + ... + P(A \cap E_n)}$$

$$= \frac{P(E_K) P(A / E_K)}{P(E_1).P(A / E_1) + P(E_2).P(A / E_2) + ... + P(E_n).P(A / E_n)}$$

Note : Baye's theorem is also known as formula for the Probability of "Causes", i.e., probability of a particular (cause) E_i given that event A has happened (already).

$P(E_i)$ is 'a priori probability' known even before the experiment, $P(A / E_i)$ "Likelihoods" and $P(E_i / A)$ 'Posteriori Probabilities' determined after the result of the experiment.

PROBLEMS

9. In a certain college, 25% of boys and 10% of girls are studying mathematics. The girls constitute 60% of the student body.

(a) What is the probability that mathematics is being studied ?

(b) If a student is selected at random and is found to be studying mathematics, find the probability that the student is a girl?

(c) a boy ?

Sol :

$$\text{Given } P(\text{Boy}) = P(B) = \frac{40}{100} = \frac{2}{5}$$

$$\text{and } P(\text{Girl}) = P(G) = \frac{60}{100} = \frac{3}{5}$$

Probability that mathematics is studied given that the student is a boy

$$= P(M / B) = \frac{25}{100} = \frac{1}{4}$$

Probability that mathematics is studied given that the student is a girl

$$= P(M / G) = \frac{10}{100} = \frac{1}{10}$$

(a) Probability that the student studied Mathematics

$$= P(M) = P(G), P(M/G) + P(B) P(M/B)$$

$\therefore$ By total probability theorem,

$$P(M) = \frac{3}{5} \cdot \frac{1}{10} + \frac{2}{5} \cdot \frac{1}{4} = \frac{4}{25}$$

(b) By Baye's theorem, probability of mathematics student is a girl

$$= P\left(\frac{G}{M}\right) = \frac{P(G)\, P(M/G)}{P(M)}$$

$$= \frac{\frac{3}{5} \cdot \frac{1}{10}}{\frac{4}{25}} = \frac{3}{8}$$

(c) Probability of maths student is a boy

$$= P\left(\frac{B}{M}\right) = \frac{P(B)\, P(M/B)}{P(M)}$$

$$= \frac{\frac{2}{5} \cdot \frac{1}{4}}{\frac{4}{25}} = \frac{5}{8}$$

10. The chance that doctor A will diagnose a disease x correctly is 60%. The chance that a patient will die by his treatment after correct diagnosis is 40% and the chance of death by wrong diagnosis is 70%. A patient of doctor A, who had disease x, died. What is the chance that his disease was diagnised correctly.

Sol :

Let E_1 be the event that "disease x is diagnosed correctly by doctor A and E_2 be the event that "a patient of doctor A who as disease x died".

$$\text{Then } P(E_1) = \frac{60}{100} = 0.6, P(E_2 / E_1)$$

$$= \frac{40}{100} = 0.4$$

Now $P(E_1) = 1 - 0.6 = 0.4$ and

$$P(E_2 / E_1) = \frac{70}{100} = 0.7$$

$\therefore$ By Baye's theorem,

$$P(E_1 / E_2) = \frac{P(E_1) \cdot P(E_1 / E_1)}{P(E_1) \cdot P(E_2 / E_1) + P(\bar{E}_1) \cdot P(E_2 / \bar{E}_1)}$$

$$= \frac{0.6 \times 0.4}{0.6 \times 0.4 + 0.4 \times 0.7} = \frac{6}{13}$$

11. A bag A contains 2 white and 3 red balls and a bag B contains 4 white and 5 red balls. One ball is drawn at random from one of the bags and it is found to be red. Find the probability that the red ball drawn is from bag B.

Sol :

Let A and B donate the events of selecting bag A and bag B respectively.

$$\text{Then } P(A) = \frac{1}{2},\ P(B) = \frac{1}{2}.$$

Let R denote the event of drawing a red ball.

Having selected bag A, the probability to draw a red ball from A $= P(R / A) = \frac{3}{5}$

Similarly $P(E / 3) = \frac{5}{9}$

One of the bags is selected at random and from it a ball is drawn at random.

It is found to be red. Ten the probability that the selected bag is B.

$$= P(B / R) = \frac{P(B) \cdot P(R / B)}{P(A) \cdot P(R / A) + P(B) P(R/ B)}$$

$$= \frac{\frac{1}{2} \cdot \frac{5}{9}}{\frac{1}{2} \cdot \frac{3}{5} + \frac{1}{2} \cdot \frac{5}{9}} = \frac{25}{52}$$

12. First box contains 2 black, 3 red, 1 white balls, second box contains 1 black, 1 red, 2 white balls and third box contains 5 black, 3 red, 4 white balls. Of these a box is selected at random. From it a red ball is randomly drawn. If the ball is red, find the probability that it is from second box.

Sol : **(Imp.)**

Let x, y, z be the first, second and third boxes.

$\therefore \quad P(x) = \frac{1}{3}, \; P(y) = \frac{1}{3}, \; P(z) = \frac{1}{3}$

Let R be the event of drawing a red ball from a box.

So $P(R / x) = \frac{3}{6} \quad P(R / y) = \frac{1}{4}, \; P(R / z) = \frac{3}{12}$

$\therefore$ By Baye's theorem, te required probability

$$= P(y/R) = \frac{P(y) \cdot P(R / y)}{P(x) P(R / x) + P(y) \cdot P(R/ y) + P(z) \cdot P(R / z)}$$

$$= \frac{\frac{1}{3} \times \frac{1}{4}}{\frac{1}{3} \times \frac{3}{6} + \frac{1}{3} \times \frac{1}{4} + \frac{1}{3} \times \frac{3}{12}} = \frac{1}{4}$$

13. In a bolt factory machines A, B, C manufacture 20%, 30% and 50% of the total of their output and 6%, 3% and 2% are defective. A bolt is drawn at random and found to be defective. Find the probabilities that it is manufactured from

(i) Machine A

(ii) Machine B

(iii) Machine C

Sol : **(Imp.)**

Let P(A), P(B), P(C) be the probabilities of the events that the bolts are manufactured by the machines A, B, C respectively. Then

$$P(A) = \frac{20}{100} = \frac{1}{5},\ P(B) = \frac{30}{100} - \frac{3}{10},\ P(C) = \frac{50}{100} = \frac{1}{2}$$

Let D denote that the bolt is defective. Then

$$P(D / A) = \frac{6}{100},\ P(D / B) = \frac{3}{100},\ \ P(D / C) = \frac{2}{100}$$

(i) If both is defective, then the probability that it is from machine A

$$= P(A / D) = \frac{P(D / A) \cdot P(A)}{P(D / A) \cdot P(A) + P(D / B) \cdot P(B) + P(D / C) \cdot P(C)} = \frac{12}{31}$$

Similarly,

(ii) $P(B / D) = \frac{9}{31}$

(iii) $P(C / D) = \frac{10}{31}$

2.5 Random Variables, Expected Values

Q9. Define Random Variables. Explain different tyeps of Random Variables.

Ans :

Introduction

We know that a variable is a quality which changes (or) varies - the change may occur due to time factor or any factor. For example, the height of a person vary with age, but the age of a person changes with time. Every variable has a range in which it can take any value.

Examples are: marks scored by a student in an examination, number of children in a family, height and weight of a person, etc. The variable could be continuous or discrete.

A continuous variable takes all possible values in its range. For example, the height of a person can take any value in a certain range, but for the sake of convenience, it is measured only up to the accuracy of inches or cms. Similarly, the weight of a person can be any value, but it is usually expressed in kgs. A discrete variable takes only certain values in a range.

For example, number of children in a family, number on a dice, etc can take only integral values.

A variable is called a random variable when there is a chance factor associated with the various values, which it can take in its range. For example, the life of a person is a random variable and there are probabilities associated with various age values.

Definition

A real variable X whose value is determined by the outcome of a random experiment is called a random variable. A random variable X can also be regarded as a real - value function defined on the sample space S of a random experiment such that for each point x of the sample space, f(x) is the probability of occurrence of the event represented by x.

Types

Random variable is of two types :

1. Discrete Random Variable
2. Continuous Random Variable

1. Discrete Random Variable

A random variable X which can take only a finite number of discrete values in an interval of domain is called a discrete random variable. In other words, if the random variable takes the values only on the set {0, 1, 2, ..., n} is called a Discrete Random variable.

Tossing of a coin, throwing a dice, the number of defectives in a sample of electric bulbs, the number of printing mistakes in each page of a book, the number of telephone calls received by the telephone operator are examples of Discrete Random variables.

Thus to each outcome 's' of a random experiment there corresponds a real number X(s) which is defined for each point of the sample S.

A few examples are :

(i) In the example (1), X(s) = {s : s = 0, 1, 2} or Range of X= {0, 1, 2}.

The random variable X is a discrete random variable.

(ii) The random variable denoting the number of students in a class is

$$X(x) = \{x : x \text{ is a positive integer}\}$$

2. Continuous Random Variable

A random variable X which can take values continuously i.e., which takes all possible values in a given interval is called a continuous random variable.

For example, the height, age and weight of individuals are examples of continuous random variable. Also temperature and time are continuous random variables.

Q10. Explain the Expected values of probability distribution.

Ans :

The behaviour of a random variable is completely characterized by the distribution function F(x) or density function f(x) or $P(x_1)$. Instead of a function, a more compact description can be made by a single numbers such as mean, median and mode known as measures of central tendency of the random variable X.

Expectation

The application of the concepts of probability theory to real life situations when the decisions are based on expectations about the value of a variable like life of an item, say electric bulb, steel, cement, scooter battery, etc.

When a dice is thrown, we know that the variable respecting the top number on the dice can be any value from 1 to 6 with probability 1/6. Now, suppose a person is not interested in listening to this statement giving the range of the variable, but all we wants to know tat single number which is expected to come up wen te dice is thrown. The answer to these types of situations when one wants te replay as a single value, is provided by the theory of expectation.

Expectation theory plays a very important role in decision - making because most of te time we take decisions based on what is expected to happen.

1. Expectation of a Discrete Variable

As defined earlier, a discrete variable takes only some finite values like number on a dice, number of children in a family, etc.

Suppose a random variable X assumes the values $x_1, x_2, ..., x_n$ with respective probability $p_1, p_2, ..l., p_n$. Then the mathematical expectation or mean or expected value of X, denoted by E(X), is defined as the sum of products of different values of x and the corresponding probabilities.

$$\therefore \quad E(X) = x_1 p_1 + x_2 p_2 + ... + x_n p_n$$

$$\text{i.e., } E(X) = \sum_{i=1}^{n} p_i x_i$$

$$\text{Similarly, } E(x^2) = \sum_{i=1}^{n} p_i . x_i^r$$

In general, the expected value of any function g(x) of a random variable X is defined as

$$E[g(x)] = \sum_{i=1}^{n} p_i \, g(x_i)$$

Note : Expected value of X is a population mean. If population mean is μ then $E(X) = \mu$.

2. Mean

The mean value μ of the discrete distribution function is given by

$$\mu = \frac{\Sigma p_i x_i}{\Sigma p_i} = \Sigma p_i x_i = E(X)$$

Note : If $E(X) = \mu$, then $E(X - \mu) = 0$.

3. Variance

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

Variance of the probability distribution of a random variable X is the mathematics expectation of $[X - E(X)]^2$. Then

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

$$\text{i.e., Var } (X) = \sum_{i=1}^{n} \left\{ [x_i - E(X)]^2 \times p(x_i) \right\}$$

4. Standard Deviation

It is the positive square root of the variance.

$$\therefore \quad S.D = \sigma = \sqrt{\sum_{i=1}^{n} p_i x_i^2 - \mu^2}$$

$$= \sqrt{E(X^2) - \mu^2}$$

$$= \sqrt{E[X - E(X)]^2}$$

Some Important Results on Variance:

(i) Variance of constant is zero i.e., V(K) = 0

(ii) If K is a constant, then $V(KX) = K^2 V(X)$

(iii) If X is a random variable and K is a constant, then V(X + K) = V(X)

(iv) If X is a discrete random variable, then $V(aX + b) = a^2 V(X)$, where V(X) is variance of X and a, b are constants.

5. If X and Y are two independent random variable, then $V(X \pm Y) = V(X) \pm V(Y)$

PROBLEMS

14. Let X denote the number of heads in a single toss of 4 fair coins. Determine (i) P(X < 2), (ii) P(I < X $\leq$ 3).

Sol :

The required probability distribution is

X	0	1	2	3	4
P(X)	$\frac{1}{16}$	$\frac{4}{16}$	$\frac{6}{16}$	$\frac{4}{16}$	$\frac{1}{16}$

(i) $P(X < 2) = P(X = 0) + P(X = 1)$

$$= \frac{1}{16} + \frac{4}{16} = \frac{5}{16}$$

(ii) $P(1 < x < 3) = P(X = 2) + P(X = 3)$

$$= \frac{6}{16} + \frac{4}{16} = \frac{10}{16} = \frac{5}{8}$$

15. Two dice are thrown. Let X assign to each point (a, b) in S the maximum of its numbers i.e., X(a, b) = max. (a, b). Find the probability distribution. X is a random variable with X(s) = {1, 2, 3, 4, 5, 6}. Also find the mean and variance of the distribution.

(OR) A random variable X has the following distribution ?

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

Find (a) the mean (b) variance (c) P(1 < x < 6)

Sol :

The total number of cases are $6 \times 6 = 36$.

The maximum number could be 1, 2, 3, 4, 5, 6 i.e., X(s) = X(a, b) – max (a, b).

The number 1 will appear only in one case (1, 1). So $p(1) = P(X - 1) = P(l, 1) = \frac{1}{36}$

For maximum 2, favourable cases are (2, 1), (2, 2), (1, 2).

So p(2) = P(X = 2) = 3/36.

For maximum 3, favourable cases are (1, 3), (3, 1), (2, 3), (3, 2), (3, 3).

So p(3) = P(X= 3) = 5/36

For maximum 4, favourable cases are (1, 4), (4, 1), (2, 4), (4, 2), (3, 4), (4, 3), (4, 4)

So p(4) = P(X = 4) = 7 / 36

Similarly, p(5) = P(X = 5)

= P((1, 5), (5, 1), (2, 5), (5, 2), (3, 5), (5, 3), (4, 5), (5, 4), (5, 5))

$= \frac{9}{36}$

p(6) = P(X= 6)

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

$= \frac{11}{36}$

$\therefore$ The required discrete probability distribution is

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

(i) Mean, $\mu = \sum_{i=1}^{6} p_i x_i = 1 \times \frac{1}{36} + 2 \times \frac{3}{36} + 3 \times \frac{5}{36} + 4 \times \frac{7}{36} + 5 \times \frac{9}{36} + 6 \times \frac{11}{36}$

$= \frac{1}{36}(1 + 6 + 15 + 28 + 45 + 46) = \frac{161}{36} = 4.47$

(ii) Variance, $\sigma^2 = \sum_{i=1}^{6} p_i x_i^2 - \mu^2$

$= \frac{1}{36}(1)^2 + \frac{3}{36}(2)^2 + \frac{5}{36}(3)^2 + \frac{7}{36}(4)^2 + \frac{9}{36}(5)^2 + \frac{11}{36}(6)^2 - (4.47)^2$

$= \frac{1}{36}(1 + 12 + 45 + 112 + 225 + 396) - (4.47)^2$

$= \frac{791}{36} - 19.9808 = 21.97 - 19.981 = 1.9912$

16. A random variable X has the following probability function :

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

(i) Determine K

(ii) Evaluate P(X < 6), P(X $\geq$ 6), P(0 < X < 5) and P(0 $\leq$ X $\leq$ 4)

(iii) if $P(X \leq K) > \frac{1}{2}$, find the minimum value of K and,

(iv) Determine the distribution function of X

(v) Mean

(vi) Variance

Sol :

(i) Since $\sum_{x=0}^{7} p(x) = 1$, we have

$K + 2K + 2K + 3K + K^2 + 2K^2 + 7K^2 + K = 1$

i.e., $10K^2 + 9K - 1 = 0$ i.e., $(10K - 1)(K + 1) = 0$

$\therefore \quad K = \frac{1}{10} = 0.1$ (since $p(x) \geq 0$, So $K \neq 1$)

(ii) $P(X < 6) = P(X = 0) + P(X = 1) + ... + P(X = 5)$

$= 0 + K + 2K + 2K + 3K + K^2 = 8K + K^2 = 0.8 + 0.01 = 0.81$ [$\because K = 0.1$]

$P(X \geq 6) = 1 - P(X < 6) = 1 - 0.81 = 0.19$

$P(0 < X < 5) = P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4)$

$= K + 2K + 2K + 3K = 8K = \frac{8}{10} = 0.8$

$P(0 \leq X \leq 4) = P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4)$

$= 0 + K + 2K + 2K + 3K = 8K = 8(0.1) = 0.8$

Note : $P(X \leq 5) = P(X = 0) + P(X = 1) + ... + P(X = 5)$

$= 0.81$ [Refer (ii)]

$P(X > 5) = 1 - P(X \leq 5) = 1 - 0.81 = 0.19$

$P(0 < X < 6) = P(X = 1) + P(X = 2) + ... + P(X = 5) = 0.81$

(iii) The required minimum value of K is obtained as below :

$P(X \leq 1) = P(X = 0) + P(X = 1) = 0 + K = \frac{1}{10} = 0.1$

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

$= \frac{1}{10} + \frac{2}{10} = \frac{3}{10} = 0.3$

$P(X \leq 3) = [P(X = 0) + P(X = 1) + P(X = 2)] + P(X = 3) = 0.3 + 0.2 = 0.5$

$P(X \leq 4) = P(X \leq 3) + P(X = 4) = 0.5 + \frac{3}{10} = 0.8 > 0.5 = \frac{1}{2}$

$\therefore$ The minimum value of K for which $P(X \leq K) > \frac{1}{2}$ is $K = 4$

(iv) The distribution function of x is given by the following table :

X	F(x) = P(X ≤ x)
0	0
1	K = 1/10
2	3K = 3/10
3	5K = 5/10
4	8K = 8/10
5	$8K + K^2 = 81/100$
6	$8K + 3K^2 = 83/100$
7	$9K + 10K^2 = 1$

(v) Mean, $\mu = \sum_{i=0}^{7} p_i x_i$

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

$= 66K^2 + 30K = \frac{66}{100} + \frac{30}{10} = 0.66 + 3 = 3.66 \quad \left(\because K = \frac{1}{10}\right)$

(vi) Variance $= \sum_{i=0}^{7} p_i x_i^2 \ \mu^2$

$= K + 8K + 18K + 48K + 25K^2 + 72K^2 + 343K^2 + 49K - (3.66)^2$

$= 440K^2 + 124K - (3.66)^2 = \frac{440}{100} + \frac{124}{10} - (3.66)^2$

$= 4.4 + 12.4 - 13.3956 = 3.4044$

17. A random variable X has the following probability distribution

Find the value of

X :	1	2	3	4	5	6	7	8
P(X)	K	2K	3K	4K	5K	6K	7K	8K

(i) K

(ii) P(X ≤ 2)

(iii) P(2 ≤ X ≤ 5)

Sol :

(i) Since $\sum_{i=1}^{n} p(x_i) = 1$, we have

$K + 2K + 3K + 4K + 5K + 6K + 7K + 8K = 1 \Rightarrow 36K = 1$

$\therefore \quad K = \frac{1}{36}$

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

$$= K + 2K = 3K = \frac{3}{36} = \frac{1}{12}$$

(iii) $P(2 \leq X \leq 5) = P(X = 2) + P(X = 3) + P(X = 4) + P(X = 5)$

$$= 2K + 3K + 4K + 5K + 14K = \frac{14}{36} = \frac{7}{18}$$

2.6 BINOMIAL DISTRIBUTION

Q11. What are the conditions under which a nominal distribution can be applied?

(OR)

What is a binomial distribution? Explain its properties and applications.

Ans : **(Imp.)**

Binomial distribution is a discrete probability distribution developed by a Swiss mathematician, 'James Bernoulli' in 1700. Thus, it is also known as Bernoulli distribution.

Properties of Binomial Distribution

The properties of Binomial distribution are as follows,

1. It describes the distribution of probabilities when there are only two mutually exclusive outcomes for each trial of an experiment for example while tossing a coin, the two possible outcomes are head and tail.
2. The process is performed under identical conditions for 'n' number of times.
3. Each trial is independent of other trials.

   It means the outcome of a particular trial does not affect the outcome of another trial.
4. The probability of success 'p' remains same for trial to trial throughout the experiment and similarly, the probability of failure (q = 1 – p) also remains constant overall the observations.
5. Binomial distribution is symmetrical when p = 0.5 [figure (i)] and it is skewed if $p \neq 0.5$, where V can be any value.

   When p > 0.5 [figure (iii)], it is skewed to the right $\rightarrow$ negatively skewed.

   When p < 0.5 [figure (ii)], it is skewed to the left -4 positively skewed. Hence, binomial distribution is 'Asymmetrical'

   When p > 0.5 and p < 0.5

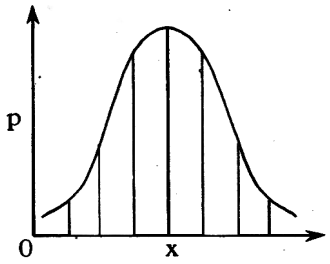


Fig. (i)

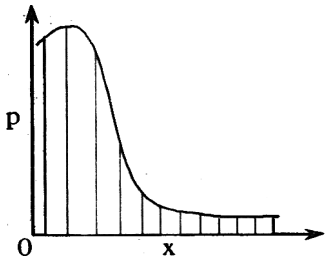


Fig. (ii)

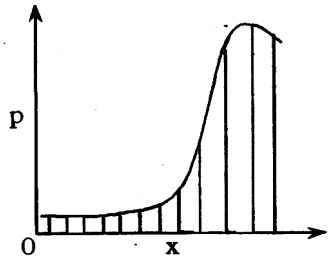


Fig. (iii)

If V is large and if neither 'p' nor '*q*' is nearly zero, in such cases the binomial distribution is modified to normal distribution by standardizing the variable, $\boxed{Z = \frac{X - np}{\sqrt{npq}}}$

Assumptions

Binomial distribution is assumed under the following conditions,

1. The number of trials V is fixed and finite.
2. Trials are independent of each other.
3. Probability of success 'p' is constant for each trial and also the probability of failure 'q' is constant for each trial.

   Always $\boxed{p + q = 1}$
4. Each trial has only two possible outcomes as success and failure.

General Form of Binomial Distribution

The general form of binomial distribution is given by the probability of obtaining exactly 'r' given number of trials 'n'.

∴ Bernoulli distribution is given by,

$$\boxed{p(r) = {}^{n}C_{r}, p^{r}q^{(n-r)}}$$

Where,

p = Probability of success in a single trial

q = Probability of failure in a single trial

p + q = 1

n = Number of trials

r = Number of successes in V number of trials.

Mean and Variance

➢ **Mean**

The mean of binomial distribution is denoted by 'p' or 'E(X)' is the expected number of successes in V number of trials.

∴ The mean of binomial distribution,

$$\boxed{\mu = np}$$

Where,

'n' = Number of trials

'p' - Probability of success in a single trial.

The two parameters of binomial distribution are 'n' and 'p'.

➢ **Variance**

The variance of binomial distribution is denoted by 'σ^2' and is the square of the standard deviation.

Thus, the variance of binomial distribution is given by,

$$\boxed{\sigma^2 = npq}$$

Where,

n - Number of trials

p = Probability of success in a single trial

q = Probability of failure in a single trial.

Standard deviation is given by,

$$\boxed{\sigma = \sqrt{npq}}$$

$$\Rightarrow \quad \boxed{\text{Standard deviation} = \sqrt{\text{Variance}}}$$

$$\Rightarrow \quad \boxed{\text{Variance, } \sigma^2 = \mu.q} \quad [\because \mu = np]$$

Applications

Binomial distribution is applicable in case of repeated trials such as,

1. Number of applications received for a junior assistant post during a particular period of time.
2. Number of births taking place in a hospital.
3. Number of candidates appearing for the screening test conducted by a company.

All the trials are statistically independent and each trial has two outcomes namely, success and failure.

Note

Binomial distribution satisfies the two essential properties of probability distribution which are explained as follows,

(i) f(x) ≥ 0

Binomial distribution fulfills this requirement as 'n' and 'p' both are positive.

Therefore, "C. (T $q^{n\,r}$ all are positive.

So $\boxed{f(x) \geq 0}$

(ii) Σf(x) = 1

Binomial expansion of $(p + q)^n$ helps in fulfilling this requirement.

As $\Sigma f(x) = \Sigma\, {}^nC_r\,,\ p^r q^{n-r}$ [∵ p = 1 – q]

$= (p + q)^r \Rightarrow (1 - q + q)^r$

$= (1)^r = 1 \Rightarrow \boxed{\Sigma f(x) = 1}$

PROBLEMS

18. Four coins were tossed 150 times and the following results were obtained

x	0	1	2	3	4
f	12	50	54	30	4

Fit binomial distribution under the assumption that the coins are unbiased

Sol :

4 coins were tossed 150 times

x	f	fx
0	12	0
1	50	50
2	54	50
3	30	90
4	4	16
	Σf = 150	**Σfx = 264**

Steps in the Fitting of Binomial Distribution

Steps 1

Calculated the Values 'p' and 'q'

In case of tossing a coin

Probability of getting a head, $p = \frac{1}{2}$

Probability of not getting a head, $q = \frac{1}{2}$

$p(r) = {}^nC_r p^r q^{n-r}$

As there are 5 terms then 'n' is one less than the number of terms

∴ n = 4, N = 150 (Σf)

Steps 2

Expand Binomial $(p + q)^n$

$(p + q)^n = \left[\frac{1}{2} + \frac{1}{2}\right]^4$

$\therefore \left[\frac{1}{2} + \frac{1}{2}\right]^4 = p(0) + p(1) + p(2) + p(3) + p(4)$

$\boxed{p(r) = {}^nC_r p^r q^{n-r}}$

n = 4

r = 0, 1, 2, 3, 4

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

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

$p(0) = {}^4C_0 \left(\frac{1}{2}\right)^0 \left(\frac{1}{2}\right)^{4-0} \quad \left[\because {}^nC_0 = 1\right]$

$= 1 \times 1 \times \left(\frac{1}{2}\right)^4 = \frac{1}{16} \quad \left[\because x^0 = 1\right]$

$= 0.0625$

$\boxed{\therefore p(0) = 0.0625}$

$p(1) = {}^4C_1 \left(\frac{1}{2}\right)^1 \left(\frac{1}{2}\right)^{4-1} \quad \left[\because {}^nC_1 = n\right]$

$= 4 \times \frac{1}{2} \times \frac{1}{8}$

$= \frac{1}{4}$

$p(1) = 0.25$

$\boxed{\therefore p(1) = 0.125}$

Step 3

Multiply each of the expected probabilities with the total frequency (N or Σf) to obtain expected frequency in each case.

$\boxed{F(r) = N \cdot P(r)}$ N = 150

F(0) = 150 × 0.0625 = 9.375

F(1) = 150 × 0.25 = 37.5

F(2) = 150 × 0.375 = 56.25

F(3) = 150 × 0.25 = 37.5

F(4) = 150 × 0.0625 = 9.375

Hence, the fitted binomial distribution is tabulated as follows.

x	f
0	9.375
1	37.5
2	56.25
3	37.5
4	9.375
	150

f = Expected frequency

19. Fit a Binomial distribution to the following data

x	0	1	2	3	4
y	28	62	46	10	4

Sol :

Steps in Fitting of Binomial Distribution

Step 1

Calculated the values 'p' and 'q'

'X' is the random variable for success and the number of Bernoulli trail here is n = 4.

[As there are 5 terms, 'n' value is always one less than the number of terms]

Mean of binomial distribution = $\frac{\Sigma fx}{\Sigma f}$

From the given data, mean can be calculated as follows,

x	f	fx
0	28	0
1	62	62
2	46	92
3	10	30
4	4	16
Total	**Σf = 150**	**Σfx = 200**

Mean = $\frac{\Sigma fx}{\Sigma f} = \frac{200}{150}$ = 1.33

$\therefore$ Mean = 1.33 .....(1)

Mean = np

From equations (1) and (2), we get,

np = 1.33 ($\because$ n = 4)

4p = 1.33

$\Rightarrow$ p = $\frac{1.33}{4}$ = 0.3325

q = 1 – 0.3325 [$\because$ q = 1 – p]

= 0.6675

$\therefore p = 0.3325 \text{ and } q = 0.6675$

Steps 2

Expand Binomial $(p + q)^n$

The binomial $(p + q)^n$ is expanded as follows,

$(p+q)^n = P(1) + p(1) + p(2) + p(3) + p(4)$

The expected binomial probabilities is given by

$\boxed{p(r) = {}^nC_r p^r q^{n-r}}$

Where,

n = 4

r = 0, 1, 2, 3, 4

p = 0.3325

q = 0.6675

$p(0) = {}^4C_0 (0.3325)^0 (0.6675)^{4-0}$ [$\because {}^nC_0 = 1$]

$= 1 \times 1 \times (0.6675)^4$ [$x^0 = 1$]

= 0.198

$\therefore p(0) = 0.198$

$p(1) = {}^4C_1(0.3325)^1 (0.6675)^{4-1}$ $[\because {}^nC_1 = n]$

$= 4 \times 0.3325 \times 0.297$

$= 0.396$

$\therefore p(1) = 0.396$

$p(2) = {}^4C_2 (0.3325)^2 (0.6675)^{4-2}$ $\left[{}^nC_r = \frac{n!}{r!(n-r)!}\right]$

$= \frac{4!}{2!(4-2)} \times 0.11 \times 0.446$

$= \frac{4\times3\times2\times1}{2\times2} \times 0.11 \times 0.446$

$= 0.2946$

$= 0.295$

$\therefore p(2) = 0.295$

$p(3) = {}^4C_3(0.3325)^3 (0.6675)^{4-3}$ $[\because {}^4C_3 = {}^4C_1 = 4]$

$= 4 \times 0.0367 \times 0.6675$ $[\because {}^nC_r = 1]$

$= 0.098$

$\therefore p(3) = 0.098$

$p(4) = {}^4C_3(0.3325)^4 (0.6675)^{4-4}$

$= 1 \times 0.0126 \times 1$ $[\because {}^nC_n = 1]$

$= 0.013$

$\therefore p(4) = 0.013$

Steps 3

Multiply each term of the expected probabilities with the total frequency (N) to obtain the expected frequencies in each case.

The expected frequencies of the binomial distribution is given by.

$F(r) = N.p(r)$ $(N = 150)$

$= 150\ p(r)$

The expected frequencies can be obtained from the table given below

Fitting of Binomial Distribution

r	p(r)	F(r) = N.p(r) = 150.p(r)
0	p(0) = 0.198	F(0) = 150 × 0.198 = 29.7
1	p(1) = 0.396	F(1) = 150 × 0.396 = 59.4
2	p(2) = 0.295	F(2) = 150 × 0.295 = 44.25
3	p(3) = 0.098	F(3) = 150 × 0.098 = 14.7
4	p(4) = 0.013	F94) = 150 × 0.013 = 1.95
	Total = 1,000	**Total = 150**

Hence, the fitted Binomial distribution is,

x	0	1	2	3	4
y	29.7	59.4	44.25	14.7	1.95

2.7 POISSON DISTRIBUTION

Q12. Explain briefly about Poisson Distribution and Applications.

Ans : **(Imp.)**

Introduction

Poisson Distribution was derived in the year 1837 by a French mathematician Simeon D Poisson. Poisson Distribution may be obtained as a limiting case of Binomial Distribution under the following conditions.

(i) n, the number of trials is infinitely large, i.e., $n \to \infty$.

(ii) p, the constant Probability of success for each trial is infinitely small, i.e., $p \to \infty$.

(iii) np = m is finite.

Under the above three conditions, the Binomial Probability Function tends to the Probability Function of the Poisson Distribution given below.

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

(or)

$$P(X = x) = P(x) = \frac{e^{-m}\ m^x}{x!} \quad x = 0, 1, 2..... \infty \quad ...(1)$$

Where,

X is the number of successes, λ = np and e = 2.71828

And $x! = x(x-1)(x-2)...3\times 2\times 1.$

Properties

Properties of Poisson Distribution are discussed in the following section.

(i) Poisson Distribution is a Discrete Probability Distribution, since the random variable X can take only values 0. 1, 2, ... ∞.

(ii) By putting r = 0,1,2,3,....., in (1), we obtain the probabilities of 0,1,2,3,.... successes respectively,

(iii) Total probability is 1.

$$\Sigma P(r) = e^{-m} + me^{-m} + (m^2/2!)\,e^{-m} + (m^3/3!)\,e^{-m} + ...$$

$$= e^{-m}\,[1 + m + m^2/2!\; + m^3/3!\; + ...]$$

$$= e^{-m} \times e^{m} = e^{-m} \times m$$

$$= e^0 = 1$$

$$E(X) = \text{Mean} = \sum_{r=0}^{\infty} rP(r) = \sum_{r=0}^{\infty} r(e^{-m})\;(m^r/r!)$$

$$= me^{-m} \times e^{m} = m$$

$$E(x^2) - [E(x)]^2 = \text{Variance}$$

$$= \sum r^2 P(r) - \left[\sum rP(r)\right]^2$$

$$= \sum r^2 P(r) - (\text{mean})^2$$

$$= [me^{-m}][e^{m}(1+m)] - m^2$$

Note : One of the special properties associated with Poisson Distribution is

Mean = Variance = m

(iv) If we know m, all the Probabilities of the Poisson Distribution can be obtained. Therefore, m is called as the parameter of the Poisson Distribution.

Applications

Some practical situations where Poisson Distribution can be used.

(i) Number of telephone calls arriving at a telephone switch board is a unit time (say, per minute)

(ii) Number of customers arriving at the super market, say, per hour.

(iii) The number of defects per unit of manufactured product (This is done for the construction of control chart for c in Statistical Quality Control)

(iv) To count the number of radio-active element per unit of time (Physics)

(v) The number of bacteria growing per unit time (Biology)

(vi) The number of defective material say, pins, blades etc. in a package manufactured by a good concern.

(vii) The number of suicides reported in a particular day.

(viii) The number of casualties (persons dying) due to rare disease such as heart attack or cancer or snake bite in a year.

(ix) Number of accidents taking place per day on a busy road.

(x) Number of typographical errors per page in a typed material or the number of printing mistakes per page in a book.

PROBLEMS

20. It is known from past experience that in a certain industrial plant there are on the average 4 industrial accidents per month. Find the probability that in a given year there will be less than 4 accidents. Assume Poisson distribution (e^{-4} = 0.0183)

Sol :

In the usual notations we are given m = 4. If the random variable X denotes the number of accidents in the industrial plant per month, then by Poisson Probability law,

$$P(X = r) = e^{-m}\, m^r/r! = e^{-4}\, 4^r/r! \qquad ...(1)$$

The required probability that there will be less than 4 accidents is given by

$P(X < 4) = [P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3)]$

$P(X < 4) = e^{-4} [1 + 4 + 4^2/2! + 4^3/3!]$

$= e^{-4} [1 + 4 + 8 + 10.67]$

$P(X < 4) = e^{-4} [23.67] = [0.0183] [23.67]$

$= 0.4332.$

---

21. If 5% of the electric bulbs manufactured by a company are defective, using Poisson distribution find the probability than in a sample of 100 bulbs : (i) none is defective, (ii) 5 bulbs are defective (Given : $e^{-5} = 0.07$).

Sol :

Here we are given n = 100,

P = Probability of a defective bulb = 5% = 0.05.

Since P is small and n is large we may approximate the given distribution by Poisson distribution. Hence the parameter m of the Poisson distribution is :

$m = np = 100 \times 0.05 = 5$

Let the random variable X denote the number of defective bulbs in a sample of 100. Then (by Poisson law)

$$P(X = r) = e^{-m} m^r / r! = e^{-5} 5^r / r! \; ; \; r = 0, 1, 2, \ldots \infty \quad \ldots(1)$$

(i) The probability that none of the bulbs is defective is given by :

$P(X = 0) = e^{-5} = 0.07$ [from 1]

(ii) The probability of 5 defective bulbs is given by :

$P(X = 5) = e^{-5} \times 5^5 / 5! = 0.07(3125/120)$

$= 0.007 (625/24) = 4.375/24 = 0.1823.$

---

22. A systematic sample of 100 pages was taken from the Concise Oxford Dictionary observed frequency distribution of foreign words per page was found to be as follows :

X	0	1	2	3	4
f(X)	123	59	14	3	1

It a Poisson Distribution to the above data.

Sol :

X = x	0	1	2	3	4	Total
f (x)	123	59	14	3	1	200
x f(x)	0	59	28	9	4	10

$$p(r) = p(X = r) = \frac{e^{-m} m^r}{r!}, \quad r = 0, 1, 2, \ldots \infty.$$

$$\bar{x} = \frac{\Sigma x f(x)}{\Sigma f(x)} = \frac{100}{200} = 0.5$$

Thus the mean (m) of the theoretical Poisson distribution is $m = \bar{x}$ = 0.5. By Poisson probability law, the theoretical frequencies are given by : f(r) = Np(r) = 200. e^{-m} $m^r/r!$); r = 0, 1, 2,

f(0) = Np(0) = 200 × e^{-m} = 200 × $e^{-0.5}$ = 200 × 0.6065 = 121.3

X	Expected Frequencies N.P(x)	Frequencies
0	NP(0)	121
1	NP(1) = N*P(0) (m)	
	= (121.3) (0.5) = 60.65	61
2	NP(2) = N*P(1) (m/2)	
	= (60.65) (0.5/2) = 15.3125	15
3	NP(3) = N*P(2)(m/3)	
	= (15.3125)(0.5/3) = 2.552	3
4	NP(4) = N *P(3) (m/4)	
	= (2.552) (0.5/4) = 0.32	0
	Total	**200**

2.8 Normal Distribution

Q13. Explain briefly about Normal Distribution.

Ans :

Introduction

The Normal Distribution was discovered by De Moivre as the limiting case of Binomial model in 1733. It was also known to Laplace no later than 1774, but through a historical error it has been credited to Gauss who first made reference to it in 1809. Throughout the 18th and 19th centuries, various efforts were made to establish the Normal model as the underlying law ruling all continuous random variables – thus the name Normal. The Normal model has, nevertheless, become the most important probability model is statistical analysis.

The normal Distribution in approximation to Binomial Distribution, whether or not p is equal to q, the Binomial Distribution tends to the form of the continuous cure when n becomes large at least for the material part of the range. As a matter of fact, the correspondence between Binomial and the Normal curve is surprisingly close even for low values of n provide dp and q are fairly near to equality. The limiting frequency curve, obtained as n, becomes large and is called the Normal frequency curve or simply the Normal curve.

Probability Density Function

A random variable X is said to have a Normal Distribution with parameters m (mean) and s^2 (Variance), if the density function is given by :

$$y = f(x) = \frac{1}{\sigma\sqrt{2x}}\, e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} ; -\infty < x < \infty$$

where y is the computed height of an ordinate at a distance of x units from the mean m.

p = $\frac{22}{7} = 3.1416$ (is a constant)

e = 2.7 183 (the base of the system of natural logarithms, and is a constant).

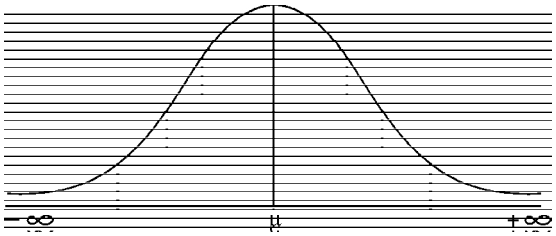
m = Mean and s^2 = Variance of the given random variable X.

In symbols, it can be expressed as :

$X \sim N(\mu, \sigma^2)$

Normal Distribution Graph

If we draw the graph of Normal Distribution, the curve obtained will be known as Normal curve and is given below :



The GRAPH of y = f(x) is a famous 'bell shaped' curve. The top of the bell is directly above the mean m. For large values of m, the curve tends to flatten out and for small values of s^2, it has a sharp peak.

When we say that curve has unit area we mean that the total area under the Normal Distribution between $(-\infty$ to $\infty)$ is equal to 1.

Standard Normal Variety (SNV)

A Random Variable with any mean and standard deviation can be transformed to a Standard Normal Variate (SNV) by subtracting the mean and dividing by the standard deviation. For a Normal Distribution with mean m and standard deviation s, the SNV 'Z' is obtained as

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

The value of Z represents the distance, expressed as a multiple of the standard deviation, that the value X lies away from the mean. The SNV Z has mean zero and variance '1'.

In symbols, if $X \sim N(\mu, \sigma^2)$, then $Z \sim N(0,1)$.

Q14. What are the properties of normal distribution ?

Ans :

The following are the important properties of the Normal Distribution :

(i) As distinguished from Binomial and Poisson Distributions where the RV is discrete, the RV associated with the Normal curve is a continuous one.

(ii) The Normal curve is symmetrical about the mean (skewness = 0). If the curve is folded along its vertical axis, the two halves will coincide.

(iii) The number of observations below the mean in a Normal Distribution are equal to the number of observations about the mean, which makes the mean and median coincide.

(iv) The height (Ordinate) of the curve for a positive deviation of 3 units from mean is same as the height of the curve for a negative deviation of 3 units from mean. This property is true for any number of units form mean or either side of it. In other words, this leads to the symmetric property.

(v) First and third quartiles are equidistant from the median.

(vi) The maximum height (Mode) of the Normal curve is at its mean. Hence the mean and mode of the Normal Distribution coincide. Thus for a Normal Distribution mean, median and mode are equal.

(vii) There is only one maximum point for the Normal which occurs at the mean. This means the Normal Distribution is a unit-modal Distribution.

(viii) The height of the curve declines as we move away on either direction from the mean. The curve approaches closure to the base (X-axis) but it never touches the X-axis. In other words, the curve is an asymptote to the base on either direction. Hence its range is unlimited or infinite on both directions.

(ix) Mean deviation about mean is (4/5) or more precisely, 0.7979 times of standard deviation.

(x) Linear combination of independent Normal variates is also a Normal variate, i.e., if X_1 and X_2 are two independent Normal variates and a_1 and a_2 are given constants, then the linear combination (a_1 X_1, a_2 X_2) will also follow a Normal distribution.

(xi) All odd moments of the Normal Distribution are zero. i.e.,

$$m_{2n+1} = 0 \text{ for } n = 0, 1, 2,....$$

(xii) $b_1 = 0$ and $b_2 = 3$ since $b_1 = 0$ the Normal Distribution is perfectly symmetrical and $b_2 = 3$ implies that Normal curve is neither leptokurtic nor platykurtic.

Limits	Area in %
$\mu \pm 0.5\ \sigma$	39.30
$\mu \pm 1.0\ \sigma$	68.26
$\mu \pm 1.5\ \sigma$	86.64
$\mu \pm 2.0\ \sigma$	95.44
$\mu \pm 2.5\ \sigma$	98.76
$\mu \pm 3.0\ \sigma$	99.74
$\mu \pm 3.5\ \sigma$	99.96

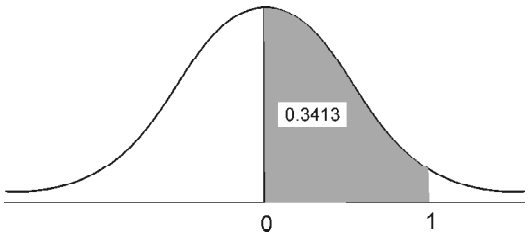
Q15. Write about area calculation – Using Normal Tables.

Ans :

This table contains the probabilities under the Normal curve between mean z = 0 and any specified positive value of z. Since the Normal curve is symmetrical, the area under the Normal curve for a negative value is same as that of the area under the Normal curve for a positive value.

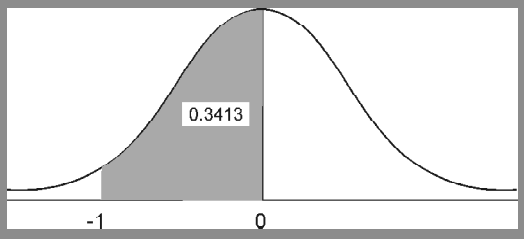
(a) Reading the area for a positive Z

For example, corresponding to z = 1, the area under the Normal curve is given as 0.3413 (from the table of the normal).



(b) Reading the area for a Negative Z

For z = –1, the area is also 0.3413 because of symmetric property of Normal Distribution.



(c) Reading the area from a Negative to positive Z

Hence,

$$Pr[-1 \le z \le +1]$$

$$= Pr[-1 \le z \le 0] + Pr[0 \le z \le +1]$$

$$= 0.3413 + 0.3413$$

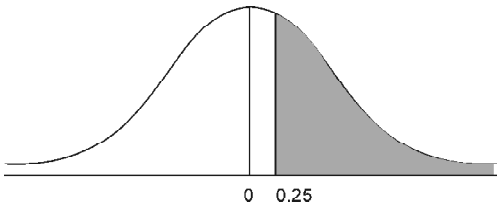
$$= 0.6826.$$

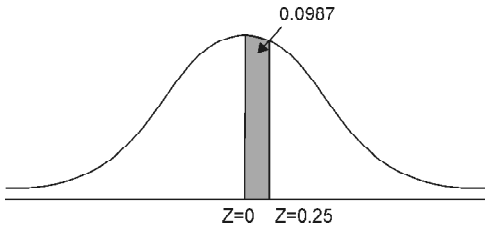
(d) Reading the area above a positive Z

If we wish to find the area under Normal curve to the right of a positive value of z we should subtract the table value from 0.5. The reason is that the Normal curve is symmetrical, the area to the right of the mean is 0.5 and the area to the right of a positive value (above a Positive value of Z) of Z is {0.5 –(The table value given for z)}.

Find the area under the Normal curve above Z = 0.25 (the right side of z = 0.25).

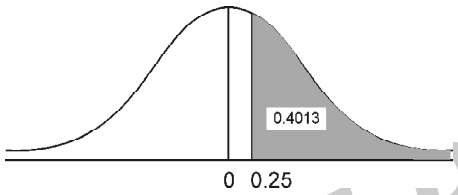
Reading this area is not possible directly from the table. To find the area required, first we read the area between Z = 0 and Z = 0.25 directly from the table as 0.0987 (the area given in the table for z = 0.25).





The required area is obtained by subtracting 0.0987 from 0.5. Hence, the area under the Normal curve above Z = 0.25 is

[0.5 – 0.0987] = 0.4013



Q16. Explain the importance of Normal distribution ?

Ans :

The Normal Distribution has great significance in statistical data analysis, because of the following reasons :

(i) The Normal Distribution has a remarkable property stated in the central limit theorem, which asserts that if X_1, X_2,.....X_n are n independent and identically random samples from a Normal Distribution which mean (m) and standard deviation (s), then the sample mean $(\bar{x})$ is also a Normal Distribution with mean (m) and standard error $\left(\frac{\sigma}{\sqrt{n}}\right)$. This result is true even if the population from which the samples are drawn is not a Normal Distribution subject to condition that n, the sample size is sufficiently large (n > 30).

(ii) Even if a variable is not Normally distributed, it can sometimes be brought to Normal form by simple transformation of variable. For example, if Distribution of X is skewed, the Distribution of $\sqrt{X}$ might come out to be Normal.

(iii) Many of the sampling Distributions like Student's t, Snedecor's F, etc. also tend to Normal Distribution.

(iv) The sampling theory and tests of significance are based upon the assumption that samples have been drawn from a Normal population with mean m and variance s^2.

(v) Normal Distribution find large applications in Statistical Quality Control.

(vi) An n becomes large, the Normal Distribution serves as a good approximation for many discrete Distributions (such as Binomial, Poisson,. etc.).

(vii) In theoretical statistics many problems can be solved only under the assumption of a Normal population. In applied work we often find that methods developed under the normal probability law yield "satisfactory" results, even when the assumption of a normal population is not fully met, despite the fact that the problem can have a normal solution only if such a premise is hypothesized.

Q17. What are the applications of Normal Distributions

Ans :

1. Production/Operations

- A workshop produces a known quantity of units per day. The average weight of unit and standard deviation are given. Assuming normal distribution, we can find how many units are expected to weigh less than greater than some given weight.
- A company manufacturers Electric bulbs and find that lifetime of the bulbs is normally distributed with some average life in hours and standard deviation in hours. On the basis of the information

it can be estimated that the number of bulbs that is expected to burn for more than specified hours and less than specified hours.

2. Finance / Accounting / Receivables

- In a business, the amount of daily collection is given for a particular period, we can estimate the average daily collection and Standard Deviation of this business. Assuming the daily collection follows a Normal Distribution.

3. Health Care and Insurance Services

- In patient's medical sample information, many parameters like cholesterol, urea in blood, hemoglobin, sugar, lipid profile, blood pressure etc., are used for diagnostic testing purpose. If each parameter follows Normal Distribution, we can calculate number of patients having abnormal and normal levels to decide upon the treatment to be followed.
- In insurance industry, if insurance Premium follows normal distribution, we can calculate how many persons fall above or below a certain insured amount to plan marketing strategies by the management.

4. Personnel

- Given with average and variance of a wage distribution of a group of workers one can estimate the number of workers in different wage ranges.
- Given a distribution of training hours of a category of employees, it can be planned for the number of hours required for training an employee to suit a particular work requirement.

PROBLEMS

23. If the salary of workers in a factory is assumed to follow a Normal Distribution with a mean of Rs. 500 and a S.D. of Rs. 100, find Number of workers whose salary vary between Rs.400 and Rs.650, give the number of workers in the factory as 15,000 ?

Sol :

The required area will be calculated only after finding the corresponding Z values as shown below.

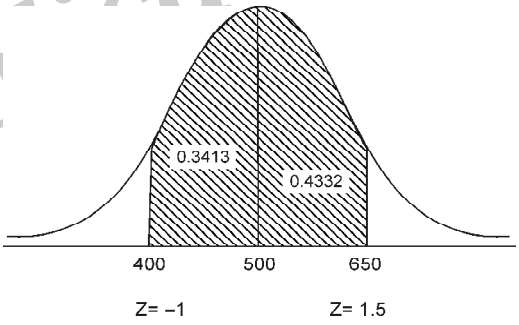
$$z_1 = \frac{x_1 - \mu}{\sigma} = \frac{(400-500)}{100} = -1$$

(left of the mean)

$$z_2 = \frac{x_2 - \mu}{\sigma} = \frac{(650-500)}{100} = +1.5$$

(right of the mean)

Now we read the area between Z = 0 to Z = 1 from table as 0.3413. Because of symmetry the area between Z = –1 to Z = 0 is same as that of the area between Z = 0 to Z = 1. Again the area between Z = 0 to Z = + 1.5 is read from table as 0.4332. Thus the desired area between is x_1 = 400 and x_2 = 650 (i.e., Z = –1 to Z = + 1.5) is (0.3413 + 0.4332 = 0.7745 as shown in the figure.



Hence, the number of workers whose salary will be between 400 and 650 is given by 0.7745 5 15,000 = 11,618.

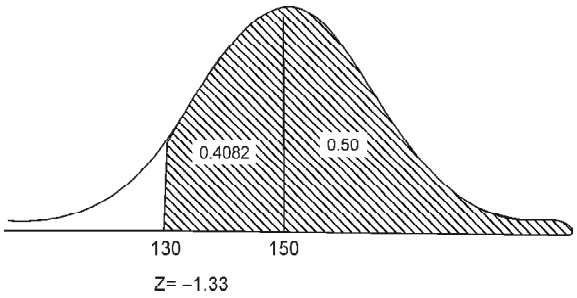
Hence, the number of workers whose salary will be between 400 and 650 is given by 0.7745 5 15,000 = 11,618.

24. A large flashlight is powered by 5 batteries. Suppose that the life of a battery is normally distributed with m = 150 hours and s = 15 hours. The flashlight will cease functioning if one or more of its batteries go dead. Assuming the lives of batteries are independent, what is the probability that flashlight will operate more than 130 hours ?

Sol : **(Imp.)**

The required area will be calculated only after finding the corresponding Z value as shown below.

$$z = \frac{x-\mu}{\sigma} = \frac{(130-150)}{15} = -1.33$$



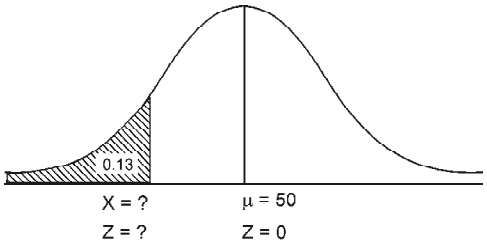
From the table we read the area from Z = 0 to Z = 1.33 as 0.4082. Due to symmetry, the area from Z = –1.33 to Z = 0 is same as 0.4082. The area to the right of mean (i.e., beyond Z = 0) is 0.5 The required area ($\geq$ 130 hours of operating time of flash light) is {0.4082 + 0.5} = 0.9082. Hence the probability that the flashlight will operate for more than 130 hours is 0.9082 and is shown in the figure below.

25. Given a normal distribution with m = 50 and s = 10, find the value of X that has (i) 13% of the area to its left and (ii) 14% of the area to its right.

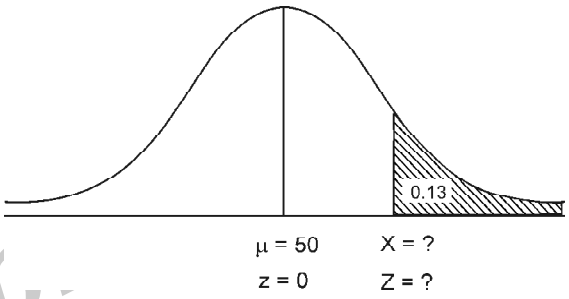
Sol :

In the previous examples, we solved fist going from a value of X to a Z value and then computing desired area. In this example, it just reverse that we begin with a known area of probability, read Z value and then determine X by rearranging the formula given below $z = \frac{x-\mu}{\sigma}$ to give X = sZ + m

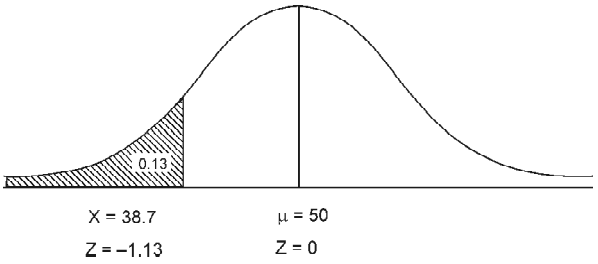
i) An area of 0.13 to the left of the desired X value is shaded in the following figure. We require a Z value that leaves an area of 0.13 to the left i.e., $P(Z < z_0) = 0.13$ as shown below.



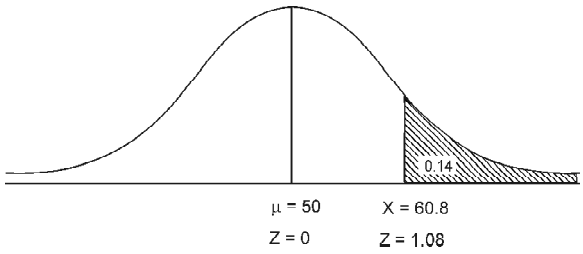
But it is not possible to read the Z value (or area) on the left side as the tables are not available for this purpose in this book. So, we use the property of symmetry and read the corresponding Z on the right side as shown below.



The table value for an area of 0.13 is 1.13. Because of symmetry this will be (–1.13) on the left side. By substituting Z = –1.13 into X = aZ + m we have X = 10(–1, 13) + 50 = 38.7. Finally, values are as given below.



ii) In this case we require a Z value that leaves 0.14 of the area to the right. This means an area of 0.46 lies between Z = 0 and a Z value to be read from the table. From the tables, we can read this value as Z = 1.08. Once again we substitute the value of Z into $X = \sigma Z + \mu$ to get X = 10(1.08) + 50 = 60.8. The final picture is as below.



2.9 Choosing Correct Probability Distribution

Q18. How to choose correct probability distribution.

Ans :

- If we plan to use a probability to describe a situation, we must be careful to choose the right one. We need to be certain that we are not using the Poisson probability distribution when it is the binomial that more nearly describes the situation we are studying.
- Remember that the binomial distribution is applied when the number of trials is fixed before the experiment begins, and each trial is independent and can result in only two mutually exclusive outcomes (success/failure, either/or, yes/no).
- Like the binomial, the Poisson distribution applies when each trial is independent. But although the probabilities in a Poisson distribution approach zero after the first few values, the number of possible values is infinite.
- The results are not limited to two mutually exclusive outcomes.
- Under some conditions, the Poisson distribution can be used as an approximation of the binomial, but not always.
- All the assumptions that form the basis of a distribution must be met if our use of that distribution is to produce meaningful results.
- We should realize that there are other useful continuous distributions.

UNIT III

Sampling and Sampling Distributions - Random sampling, Non-Random Sampling distributions, operational considerations in sampling.

Estimation - Point estimates, interval estimates, confidence intervals, calculating interval estimates of the mean and proportion, t-distribution, determination of sample size in estimation.

3.1 SAMPLING

Q1. Define sampling.

(OR)

Explain the term Sampling.

Ans :

Meaning

Sampling may be defined as the selection of some part of an aggregate or totality on the basis of which a judgement or inference about the aggregate or totality is made. In other words, it is the process of obtaining information about an entire population by examining only a part of it.

In most of the research work and surveys, the usual approach happens to be to 'make generalizations' or to 'draw inferences' based on samples about the parameters of population from which the samples are taken. The researcher quite often selects only a few items from the universe for his study purpose. All this is done on the assumption that the sample data will enable him to estimate the population parameters.

The items so selected constitute what is technically called a sample, their selection process is called sample design and the survey conducted on the basis of sample is described a sample survey. Sample should be truly representative of population characteristics without any bias so that it may result in valid and reliable conclusions.

Q2. Explain the basic terminology are used in Sampling.

(OR)

State the Basic terminology used in sampling.

Ans : **(Imp.)**

1. Population

Population is a collection of all objects, animate or inanimate under study is known as a population. It can be a finite population with a finite number of objects and finite population with infinite number of objects.

Example

(i) The numbers of students in a class is 50. Here N = 50 which is finite i.e., the population is finite.

(ii) The number of stars in the sky. Here N is infinite. Therefore the population is infinite population.

2. Sample

A finite subset of the population is called as sample. It is also a part of the population.

Sample Size : is the number of objects in the sample.

Example

Selecting 2 girls in a class of 20 students is a sample

Note

A sample with less than 30 items is a small sample and a sample with more than 30 items is a large sample.

3. Random Sample

A random sample is one in which each member of the population has equal chance or probability of being included in the sample which is taken from the population.

4. Parameters of Statistics

(i) Population parameters or parameters

The measures of population are mean (μ), variance (σ^2), standard deviation (σ) and proportions (P).

(ii) Sample statistics or statistics

The measures computed from the sample observations are mean ($\bar{X}$), variance (s^2) standard deviation (S) and proportions (P).

Q3. What is the purpose (or) importance of sampling ?

Ans :

- The importance of sampling is that you can determine the adequate respondents from the total number of target population.
- Thus, it will be used in the research study which should be adequate to warrant generalization of the findings to the target population.
- The sample size represents the characteristics of the whole population (representativeness of the sample).
- The sampling are it is economical and practical faster and cheaper it can yield more comprehensive information
- It is more accurate and because of savings it permits in time and money, the sample survey makes possible the use of much larger and much more varied populations than would be possible for the same expenditure if one were making a complete enumeration.

Q4. What are the advantages of sampling?

(OR)

State the advantages of sampling.

Ans :

Sampling ensures convenience, collection of intensive and exhaustive data, suitability in limited resources and better rapport. In addition to this, sampling has the following advantages also.

1. Low cost of sampling

If data were to be collected for the entire population, the cost will be quite high. A sample is a small proportion of a population. So, the cost will be lower if data is collected for a sample of population which is a big advantage.

2. Less time consuming in sampling

Use of sampling takes less time also. It consumes less time than census technique. Tabulation, analysis etc., take much less time in the case of a sample than in the case of a population.

3. Scope of sampling is high

The investigator is concerned with the generalization of data. To study a whole population in order to arrive at generalizations would be impractical.

Some populations are so large that their characteristics could not be measured. Before the measurement has been completed, the population would have changed. But the process of sampling makes it possible to arrive at generalizations by studying the variables within a relatively small proportion of the population.

4. Accuracy of data is high

Having drawn a sample and computed the desired descriptive statistics, it is possible to determine the stability of the obtained sample value. A sample represents the population from which its is drawn. It permits a high degree of accuracy due to a limited area of operations. Moreover, careful execution of field work is possible. Ultimately, the results of sampling studies turn out to be sufficiently accurate.

5. Organization of convenience

Organizational problems involved in sampling are very few. Since sample is of a small size, vast facilities are not required. Sampling is therefore economical in respect of resources. Study of samples involves less space and equipment.

6. Intensive and exhaustive data

In sample studies, measurements or observations are made of a limited number. So, intensive and exhaustive data are collected.

7. Suitable in limited resources

The resources available within an organization may be limited. Studying the entire universe is not viable. The population can be satisfactorily covered through sampling. Where limited resources exist, use of sampling is an appropriate strategy while conducting marketing research.

3.1.1 Random Sampling, Non-Random Sampling Distributions

Q5. Explain different types of sampling methods.

Ans : **(Imp.)**

Some important methods of sampling are discussed below :

(I) Probability Sampling Methods

1. Random Sampling (or) Probability Sampling

It is the process of drawing a sample from a population in such a way that each member of the population has an equal chance of being included in the sample. The sample obtained by the process of random sampling is called a random sample.

For example :

(i) A handy of cards dealt from a well - shuffled pack of cards is a random sample.

(ii) Selecting randomly 20 words from a dictionary is a random sample.

(iii) Choosing 10 patients from a hospital in order to test the efficacy of a certain newly - invented drug.

If each element of a population may be selected more than once then it is called sampling with replacement whereas if the element cannot be selected more than once, it is called sampling without replacement.

Note :

If N is the size of a population and n is the sample size, then

(i) The number of samples with replacement $= N^n$

(ii) The number of samples without replacement $= {}^NC_n$

2. Stratified Sampling (or) Stratified Random Sampling

This method is useful when the population is heterogeneous. In this type of sampling, the population is first sub-divided into several parts (or small groups) called strata according to some relevant characteristics so that each stratum is more or less homogeneous. Each stratum is called a sub-population. Then a small sample (called sub-sample) is selected from each stratum at random. All the sub-samples are combined together to form the stratified sample which represents the population properly. The process of obtaining and examining a stratified sample with a view to estimating the characteristic of the population is known as Stratified Sampling.

For example, let us select a stratified sample of 500 families from a city having 50,000 families, with a view to studying their economic condition. For this purpose, the city area is divided into a number of strata, according to economic condition of their inhabitants, as measured by annual income (say). Thus, localities mostly inhabited by people with more or less similar annual income may be included under one stratum. A few families are then chosen at random from each so that the sum total of all the families from all the strata is 500.

3. Systematic Sampling (or) Quasi - Random Sampling

As the name suggests this means forming the sample in some systematic manner by taking items at regular intervals. In this method, all the units of the population are arranged in some order. If the population size is finite, all the units of the population are arranged in some order. Then from the first k items, one unit is selected at random. This unit and every the unit of the serially listed population combined together constitute a systematic sample. This type of sampling is known as Systematic Sampling.

The difference between random sampling and systematic sampling lies in the fact that in the case of a random sample all the members have to be chosen randomly, whereas in the case of a systematic sample only the first member has to be chosen at random.

II. Non-Probability Sampling Methods.

1. Purposive Sampling (or) Judgement Sampling

When the choice of the individual items of a sample entirely depends on the individual judgement of the investigator (or sampler), it is called a Purposive or Judgement Sampling.

In this method, the members constituting the sample are chosen not according to some definite scientific procedure, but according to convenience and personal choice of the individual, who selects the sample. Two or more such independent purposive samples, may give widely different estimates of the same population. In this type, the investigator must have a good deal of experience and a thorough knowledge of the population. Purposive selection is always subject to some kind of bias. This method is suitable when the sample is small.

For example, if a sample of 20 students is to be selected from a class of 100 to analyse the extra-curricular activities of the students, the investigator would select the students who, in his judgement, would represent the class.

2. Sequential Sampling

It consists of a sequence of sample drawn one after another from the population depending on the results of previous samples. If the result of the first sample leads to decision which is not acceptable, the lot from which the sample was drawn is rejected. But if the result of the first sample is acceptable, no new sample is drawn. But if the first sample leads to no clear decision, a second sample is drawn and, as before, if required third sample is drawn to arrive at a final decision to accept or reject the lot. It is widely used in Statistical Quality Control in factories engaged in mass production and other areas.

3.2 Operational Considerations in Sampling

Q6. What are the Operational Considerations in Sampling

Ans :

The standard error, σ_x is a measure of dispersion of the sample means around the population mean. If the dispersion decreases (if σ_x becomes smaller), then the values taken by the sample mean tend to cluster more closely around μ. Conversely, if the dispersion increases (if σ_x becomes larger), the values taken by the sample mean tend to cluster less closely around μ. We can think of this relationship this way: As the standard error decreases, the value of any sample mean will probably be closer to the value of the population mean. Statisticians describe this phenomenon in another way: As the standard error decreases, the precision with which the sample mean can be used to estimate the population mean increases.

The Finite Population Multiplier

To this point in our discussion of sampling distributions, we have used Equation to calculate the standard error of the mean :

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

This equation is designed for situations in which the population is infinite, or in which we sample from a finite population with replacement (that is, after each item is sampled, it is put back into the population before the next item is chosen.

Many of the populations decision makers examine are finite, that is, of stated or limited size.

Standard Error of the Mean for Finite Populations

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \times \sqrt{\frac{N-n}{N-1}}$$

where

- N = size of the population
- n = size of the sample

This new term on the right - hand side, which we multiply by our original standard error, is called the finite population multiplier :

Finite Population Multiplier

$$\text{Finite population multiplier} = \sqrt{\frac{N-n}{N-1}}$$

3.3 SAMPLING DISTRIBUTIONS

Q7. What is Sampling Distributions? Explain the characteristics of Sampling Distributions.

Ans : **(Imp.)**

Meaning

Samples of a given size may be drawn randomly from a population and the statistical constants like mean, standard deviation may be computed for each such sample. The distribution of each such statistic is called sampling distribution.

Characteristics

- Sampling distribution is a frequency distribution representing the means taken from a great many samples, of the same size.
- The main characteristic of this is that it approaches normal distribution even went the population distribution is not normal provided the sample size is sufficiently large (greater than 30).
- The significance of the sampling distribution follows from the fact that the mean of sampling distribution is the same as the mean of the population.

Q8. Define sample size determination.

Ans :

When using a sampling method, size of the sample has to be determined, various experts have different views. The sample size depends on various factors. It could be the time limits, the financial aspect, accuracy point to be achieved. Sampling theory is not of much help to determine size of sample.

There are two considerations to determine sample size. They are,

(i) Sample size should be directly proportional to the variation in the individual item. As sample size increases as variation in individual item increases and vice versa.

(ii) Larger the sample size, H_0 is the accuracy i.e., sample size $\propto$ accuracy level.

3.4 ESTIMATION

Q9. Define the term Estimation.

Ans :

- When data are collected by sampling from a population, the most important objective of statistical analysis is to draw inferences (or) generalizations about that population from the information embodied in the sample.
- Statistical estimation, or briefly estimation, is concerned with the methods by which population characteristics are estimated from sample information. It may be pointed out that the true value of a parameter is an unknown constant that can be correctly ascertained only by an exhaustive study of the population.
- However, it is ordinarily too expensive or it is infeasible to enumerate complete populations to obtain the required information.
- In case of finite populations, the cost of complete censuses may be prohibitive.
- In case of infinite population, complete enumerations are impossible.
- A realistic objective may be to obtain a guess or estimate of the unknown true value or an interval of values from the sample data and also to determine the accuracy of the procedure.
- Statistical estimation procedures provide us with the means of obtaining estimates of population parameters with desired degrees of precision.

3.4.1 Types

3.4.1.1 Point estimates, interval estimates

Q10. Explain briefly about various types of estimates.

Ans : **(Imp.)**

1. Point Estimation

If an estimate of the population parameter is given by a single value, then the estimate is called a Point Estimation of the parameter. But if an estimate of a population parameter is given by two different values between which the parameter may be considered to lie, then the estimate is called an interval estimation of the parameter.

Example : If the height of a student is measured as 162 cms, then the measurement gives a point estimation. But if the height is given as (163 $\pm$ 3.5) cms, then the height lies between 159.5 cms and 166.5 cms and the measurement gives an interval estimation.

The sample mean $\bar{x}$ is a point estimate of population mean μ, sample variance s^2 is a point estimate of population variance σ^2.

Definition

A point estimate of a parameter 0 is a single numerical value, which is computed from a given sample and serves as an approximation of the unknown exact value of the parameter.

A point estimator is a statistic for estimating the population parameter θ and will be denoted by $\hat{\theta}$ (read as theta hat).

Properties of Estimation

(i) Unbiased Estimator

Let $\hat{\theta}$ be an estimator of θ. The statistic θ is said to be an unbiased estimator, or its value an unbiased estimate, if, and only if the mean or expected value of $\hat{\theta}$ is equal to θ. This is equivalent to say that the mean of the probability distribution of $\hat{\theta}$ (or the mean of the sampling distribution of $\hat{\theta}$) is equal to θ. An estimator possessing this property is said to be unbiased.

(ii) Unbiased estimator

A statistic or point estimator θ is said to be an unbiased estimator of the parameter θ if E ($\hat{\theta}$) = θ. In other words, if E (statistic) = parameter, then statistic is said to be an unbiased estimator of the parameter.

(iii) Variance of a point estimator

If $\hat{\theta}_1$ and $\hat{\theta}_2$ are two unbiased estimators of the same population parameter θ, we would choose the estimator whose sampling distribution has the smaller variance. Hence, if $\sigma^2_{\hat{\theta}_1} < \sigma^2_{\hat{\theta}_2}$, we say that $\hat{\theta}_1$ is a more efficient estimator of θ than $\hat{\theta}_2$.

2. Interval Estimation

Point estimates rarely coincide with quantities they are intended to estimate. So instead of point estimation where the quantity to be estimated is replaced by a single value a better way of estimation is interval estimation, which determines an interval in which the parameter lies.

Even the most efficient unbiased estimator cannot estimate the population parameter exactly. It is true that our accuracy increases with large samples. But there is still no reason why we should expect a point estimate from a given sample to be exactly equal to the population parameter, it is supposed to estimate.

Therefore in many situations it is preferable to determine an interval within which we would expect to find the value of the parameter. Such an interval is called an interval estimate. Thus an interval estimate is an interval (confidence interval) obtained from a sample.

An interval estimate of a population parameter θ is an interval of the form $\hat{\theta}_L < \theta < \hat{\theta}_u$ where $\hat{\theta}_L$ and $\hat{\theta}_u$ depend on the value of the statistic $\hat{\theta}$ for a particular sample and also on the sampling distribution of $\hat{\theta}$.

Since different samples will generally yield different values of $\hat{\theta}$ and, therefore, different values $\hat{\theta}_L$ and $\hat{\theta}_u$. These end points of the interval are values of corresponding random variables $\hat{\theta}_L$ and $\hat{\theta}_u$.

From the sampling distribution of $\hat{\theta}$ we shall be able to determine $\hat{\theta}_L$ and $\hat{\theta}_u$ such that the $P(\hat{\theta}_L < \theta < \hat{\theta}_u)$ is equal to any positive fractional value we care to specify. If, for instance, we find $\hat{\theta}_L$ and $\hat{\theta}_u$ such that $f(\hat{\theta}_L < \theta < \hat{\theta}_u) = 1 - \alpha$ for $\theta < \alpha < 1$, then we have a probability of $1 - \alpha$ of selecting a random sample that will produce an interval containing 0. The interval $\hat{\theta}_L < \theta < \hat{\theta}_u$ computed from the selected sample, is then called a $(1 - \alpha)$ 100% confidence interval. The fraction $1 - \alpha$ called the confidence coefficient or the degree of confidence, and the end points, $\hat{\theta}_L$ and $\hat{\theta}_u$, are called the lower and upper confidence limits.

Thus, when $\alpha = 0.05$, we have a 95% confidence interval, and when $\alpha = 0.01$, we obtain a wider 99% confidence interval.

Q11. What are the properties of a good estimator?

Ans : **(Imp.)**

A good estimator, as common sense dictates, is close to the parameter being estimated. Its quality is to be evaluated in terms of the following properties

(i) Unbiasedness

An estimator is said to be unbiased if its expected value is identical with the population parameter being estimated. That is if $\hat{\theta}$ is an unbiased estimate of 0, then we must have $E(\hat{\theta}) = 0$. Many estimators are "A symptomatically unbiased" in the sense that the biases reduce to practically insignificant values zero when n becomes sufficiently large. The estimator S^2 is an example.

It should be noted that bias in estimation is not necessarily undesirable. It may turn out to be an asset in some situations. For example, it may happen that an unbiased estimator is less desirable than a biased estimator if the former has a greater variability than the latter and, as a consequence, the expected value of the latter is closer than that of the former to the parameter being estimated.

(ii) Consistency

If an estimator, say $\hat{\theta}$ approaches the parameter θ closer and closer as the sample size n increases, $\hat{\theta}$ is said to be a consistent estimator of G. Stating somewhat more regorously, the estimator $\hat{\theta}$ is said to be a consistent estimator of θ if, as n approaches infinity, the probability approaches 1 that $\hat{\theta}$ will differ from the parameter θ by not more than an arbitrary small constant.

The sample mean is an unbiased estimator of p no matter what form the population distribution assumes, while the sample median is an unbiased estimate of μ only if the population distribution is symmetrical. The sample mean is better than the sample median as an estimate of μ in terms of both unbiasedness and consistency.

In case of large samples consistency is a desirable property for an estimator to possess. However in small samples, consistency is of little importance unless the limit of probability defining consistency is reached even with a relatively small size of the simple.

(iii) Efficiency

The concept of efficiency refers to the sampling variability of an estimator. If two competing estimators are both unbiased, the one with the smaller variance (for a given sample size) is said to be relatively more efficient. Stated in a somewhat different language, estimator $\hat{\theta}_1$ is said to be more efficient than another estimator $\hat{\theta}_2$ for θ if the variance of the first is less than the variance of the second. The smaller the variance of the estimator, the more concentrated is the distribution of the estimator around the parameter being estimated and, therefore, the better this estimator is.

If the population is symmetrically distributed, then both the sample mean and the sample median are consistent and unbiased estimators of μ. Yet the sample mean is better than the sample median as an estimator of θ. This claim is made in terms of efficiency.

(iv) Sufficiency

An estimator is said to be sufficient if it conveys as much information as is possible about the parameter which is contained in the sample. The significance of sufficiency lies in the fact that if a sufficient estimator exists, it is absolutely unnecessary to consider any other estimator; a sufficient estimator ensures that all information a sample can furnish with respect to the estimation of a parameter is being utilized.

Many methods have been devised for estimating parameters that may provide estimators satisfying these properties. The two important methods are the least square method and the method of maximum likelihood.

Having discussed the above concepts let us now discuss the various situations where we have to apply different tests of significance. For the sake of convinance and clarity these situations may be summed up under the following three heads :

- Test of Significance for Attributes.
- Tests of Significance for Variables (Large Samples).
- Tests of Significance for Variables (Small Samples).

3.4.2 Confidence Intervals

Q12. State the various Confidence Intervals of estimator?

Ans :

A Confidence Interval is a range of values we are fairly sure our true value lies in.

Example: Average Height

We measure the heights of 40 randomly chosen men, and get a mean height of 175cm,

We also know the standard deviation of men's heights is 20cm.

The 95% Confidence Interval

175 cm ± 6.2 cm

This says the true mean of ALL men (if we could measure all their heights) is likely to be between 168.8 cm and 181.2 cm.

The "95%" says that 95% of experiments like we just did will include the true mean, but 5% won't.

So there is a 1-in-20 chance (5%) that our Confidence Interval does Not include the true mean.

Calculating the Confidence Interval

Step 1: find the number of observations n, calculate their mean X, and standard deviations

Using our example:

- Number of observations: n = 40
- Mean: X = 175
- Standard Deviation: s = 20

Note:

we should use the standard deviation of the entire population, but in many cases we won't know it.

We can use the standard deviation for the sample if we have enough observations (at least n=30, hopefully more).

Step 2:

decide what Confidence Interval we want: 95% or 99% are common choices. Then find the "Z" value for that Confidence Interval here:

Confidence Interval	Z
80%	1.282
85%	1.440
90%	1.645
95%	1.960
99%	2.576
99.5%	2.807
99.9%	3.291

For 95% the Z value is 1.960

Step 3:

use that Z in this formula for the Confidence Interval

$$X \pm Z\, s\sqrt{n}$$

Where:

- X is the mean
- Z is the chosen Z-value from the table above
- s is the standard deviation
- n is the number of observations

And we have:

$175 \pm 1.960 \times 20 - 40$

Which is:

175 cm ± 6.20 cm

In other words : from 168.8 cm to 181.2 cm

The value after the ± is called the margin of error

The margin of error in our example is 6.20 cm

3.4.3 Calculating Interval Estimates of the mean and Proportion

Q13. Explain the Interval Estimates of the mean and Proportion.

Ans :

I. Confidence limits for Population Mean μ

(i) 95% confidence limits are $\bar{x}$ ± l .96 (S.E. of $\bar{x}$)

(ii) 99% confidence limits are $\bar{x}$ ± 2.58 (S.E. of $\bar{x}$)

(iii) 99.73% confidence limits are $\bar{x}$ ± 3 (S.E. of $\bar{x}$)

(iv) 90% confidence limits are $\bar{x}$ ± 1.64 (S.E. of $\bar{x}$)

II. Confidence limits for Population Proportion P

(i) 95% confidence limits are p ± 1.96 (S.E. of p)

(ii) 99% confidence limits are p ± 2.58 (S.E. of p)

(iii) 99.73% confidence limits are p ± 3 (S.E. of p)

(iv) 90% confidence limits are p ± 1.64 (S. E. of p)

III. Confidence limits for the difference $\mu_1 - \mu_2$ of two Population Means μ_1 and μ_2

(i) 95% confidence limits are $(\bar{x}_1 - x_2)$ ± 1.96 (S.E. of $(\bar{x}_1 - \bar{x}_1)$)

(ii) 99% confidence limits are $(\bar{x}_1 - x_2)$ ± 2.58(S.E. of $(\bar{x}_1 - \bar{x}_1)$)

(iii) 99.73% confidence limits are $(\bar{x}_1 - x_2)$ ± 3 (S.E. of $(\bar{x}_1 - \bar{x}_1)$)

(iv) 90% confidence limits are $(\bar{x}_1 - x_2)$ ± 1.64(S.E. of $(\bar{x}_1 - \bar{x}_1)$)

IV. Confidence limits for the difference $P_1 - P_2$ of two population proportions

(i) 95% confidence limits are $(p_1 - p_2) \pm 1.96$ (S.E. of $(p_1 - p_2)$)

(ii) 99% confidence limits are $(p_1 - p_2) \pm 2.58$ (S.E. of $(p_1 - p_2)$)

(iii) 99.73% confidence limits are $(p_1 - p_2) \pm 3$ (S.E. of $(p_1 - p_2)$)

(iv) 90% confidence limits are $(p_1 - p_2) \pm 1.64$ (S.E. of $(p_1 - p_2)$)

Note :

If for a statistic, $P(-3 < z < 3)$ i.e., confidence limits cover 99.73% of the area under the standard normal curve, then we say that the confidence limits for the statistic are almost sure limits without mentioning the degree of confidence.

In calculating S.E. if population parameters are unknown, corresponding sample statistics are used to find an approximate value of S. E. For example, if the population S. D., σ. is not given, the S.E. of $\bar{x}$ is calculated by using sample S.D. s in place of σ.

3.4.4 Determination of Sample Size in Estimation

Q14. Explain the determination of Sample Size in Estimation.

Ans :

Determination of proper sample size is important for testing of hypothesis in business problems. The size of sample should neither be too small nor too large. If the size of the sample is too small, then it may not give a valid conclusion. On the other hand, if the size of the sample is too large, then there may be loss of time and money without getting the required results.

1. Sample size for estimating population Mean

Let x be the mean of a random sample drawn from a population having mean μ and S.D. a. Let the sampling distribution of the sample mean $\bar{x}$ be approximately a normal distribution with mean p and S, D. σ. If E be the permissible sampling error, then

$$E = \bar{x} - \mu$$

The confidence interval for the population mean μ is $\bar{x} \pm z$(S.E. of $\bar{x}$) = $\bar{x} \pm E$

where z = confidence coefficient or z – value (which is 1.96 at 5% level of significance) and

$E = z$ (S.E. of $\bar{x}$) = $z.\frac{\sigma}{\sqrt{n}}$, n being the sample size and σ = population S.D.

Thus $E = \frac{2\sigma}{\sqrt{n}}$ or $\sqrt{n} = \frac{z\sigma}{E}$

$\therefore n = \left(\frac{z\sigma}{E}\right)^2$ the required sample size for estimating population mean.

2. Sample size for Estimating Population Proportion

Let p be the sample proportion and P be the population proportion. If E be the permissible sampling error, then E = p – P and the confidence interval for the population proportion P is $p \pm z$ (S.E. of p) = p $\pm$ E where $E = z$ (S.E. of p) = $z\sqrt{\frac{PQ}{n}}$, n being the sample size, or, $E^2 = \frac{z^2PQ}{n}$ or $\frac{z^2PQ}{E^2}$ where Q = 1P.

Proposition

If $x_i, x_2,....,x_n$ be a random sample from an infinite population with variance σ^2 and sample mean $\bar{x}$, then — then $\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ is a biased estimate of σ^2, but $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ is an unbiased estimate of σ^2.

3. Maximum Error of Estimate E for Large Samples

Since the sample mean estimate very rarely equals to the mean of population μ, a point estimate is generally accompanied with a statement of error which gives difference between estimate and the quantity to be estimated, the estimator.

Thus error is $|\bar{x} - \mu|$

For large n, the random variable $\frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$ is normal variate approximately.

Then $P(-z_{a/2} < z < z_{a/2}) = 1 - \alpha$ where $z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$

Hence $P\left(-z_{\alpha/2} < \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} < z_{\alpha/2}\right) = 1 - \alpha$

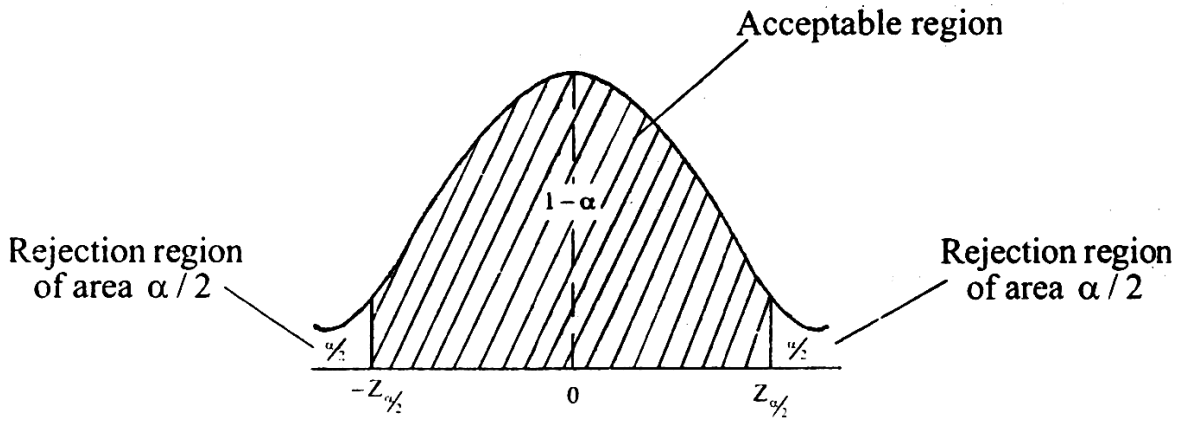


Fig.: $P(-z_{\alpha/2} < z_{\alpha/2}) = 1 - \alpha$

Multiplying each term in the inequality by a/yfn , and then subtracting $\bar{x}$ from each term and multiplying by – 1.

$$\therefore \quad P\left(\bar{x} - z_{\alpha/2}\cdot\left(\frac{\sigma}{\sqrt{n}}\right) < \mu < \bar{x} + z_{\alpha/2}\cdot\left(\frac{\sigma}{\sqrt{n}}\right)\right) = 1 - \alpha$$

Confidence interval for μ,σ known :

If $\bar{x}$ is the mean of a random sample of size n from the population with known variance σ^2, $(1 - \alpha)$

$$\bar{x} - z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) < m < \bar{x} + z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$$

where $z_{\alpha/2}$ is the z-value leaving an area of $\alpha/2$ to the right.

So, the maximum error of estimate E with $(1 - \alpha)$ probability is given by

$$E = z_{\alpha/2}.\left(\frac{\sigma}{\sqrt{n}}\right)$$

Thus in the point estimation of population mean p with sample mean $\bar{x}$ for a large random sample ($n \geq 30$), one can assert with probability $(1 - \alpha)$ that the error

$|\bar{x} - \mu|$ will not exceed $z_{\alpha/2}.\left(\frac{\sigma}{\sqrt{n}}\right)$.

Sample size :

When α, E, σ are known, the sample size n is given by

$$n = \left(\frac{z_{\alpha/2}\sigma}{E}\right)^2$$

When CT is unknown : In this case, σ is replaced by s, the standard deviation of sample to determine E.

Thus the maximum error estimate $E = t_{\alpha/2}.\frac{s}{\sqrt{n}}$ = with (1– a) probability.

4. Maximum Error of Estimate E for Small Samples

When $n < 30$, small sample, we use S, the S.D of sample to determine E. When a is not known, t can be used to construct a confidence interval as μ.

The procedure is the same as that with known a except that a is replaced by S and the standard normal distribution is replaced by the t-distribution.

$P(-t_{\alpha/2} < t < t_{\alpha/2}) = 1 - \alpha$

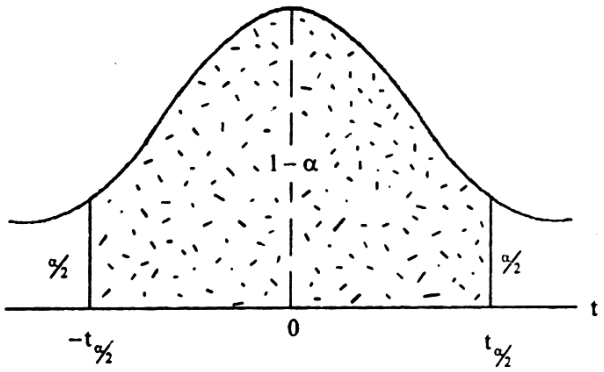


Fig.: $P(-t_{\alpha/2} < t < t_{\alpha/2}) = 1 - \alpha$

Where $t_{\alpha/2}$ is the t-values within (n – 1) degrees of freedom above which we find an area of $\alpha/2$. Because of symmetry, an equal of $\alpha/2$ will fall to the left of $-r_{\alpha/2}$.

Substituting for t, we write $P\left(-t_{\alpha/2} < \frac{\bar{x}-\mu}{S/\sqrt{n}} < t_{\alpha/2}\right) = 1 - \alpha$

Multiplying each term in the inequality by $S/\sqrt{n}$, and then subtracting $\bar{x}$ from each term and multiplying by – 1, we obtain.

$$P(\bar{x} - t_{\alpha/2}(S/\sqrt{n}) < \mu < \bar{x} + t_{\alpha/2}(S/\sqrt{n})) = 1 - \alpha.$$

PROBLEMS

1. In a study of an automobile insurance a random sample of 80 body repair costs had a mean of Rs. 472.36 and the S.D of Rs. 62.35. If $\bar{x}$ is used as a point estimate to the true average repair costs, with confidence we can assert that the maximum error doesn't exceed Rs. 10.

Sol :

Size of a random sample, n = 80

The mean of random sample, $\bar{x}$ = Rs. 472.36

Standard deviation = 62.35

Maximum error of estimate, E_{max} = Rs. 10

We have $E_{max} = z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \Rightarrow z_{\alpha/2} = \frac{E_{max} \cdot \sqrt{n}}{\sigma} = \frac{10\sqrt{80}}{62.35} = \frac{89.4427}{62.35} = 1.4345$

$\Rightarrow \quad z_{\alpha/2} = 1.43$

The area when z = 1.43 from table is 0.4236.

$\therefore$ Confidence = $(1 - \alpha)$ 100% = 84.72%

Hence we are 84.72% confidence that the maximum error is Rs. 10.

2. Assuming that σ = 20.0, how large a random sample be taken to assert with probability 0.95 that the sample mean will not differ from the true mean by more that 3.0 points ?

Sol :

Given maximum error E = 3.0 and σ = 20.0

We have $z_{\alpha/2} = 1.96$

We know that, $n = \left(\frac{z_{\alpha/2} \cdot \sigma}{E}\right)^2$

$\Rightarrow \quad n = \left(\frac{1.96 \times 20}{3}\right)^2 = 170.74$

$\therefore \quad n \cong 171$

3. It is desired to estimate the mean number of hours of continuous use until a certain computer will first require repairs. If it can be assumed that σ = 48 hours, how large a sample be needed so that one will be able to assert with 90% confidence that the sample mean a off by at most 10 hours.

Sol :

It is given that

Maximum error, E = 10 hours

σ = 48 hours

and $z_{\alpha/2}$ = 1.645 (for 90%)

$$\therefore n = \left(\frac{z_{\alpha/2} \cdot \sigma^2}{E}\right)^2 = \left(\frac{1.645 \times 48}{10}\right)^2 = 62.3 = 62$$

Hence sample size = 62

4. What is the maximum error one can expect to make with probability 0.90 when using the mean of a random sample of size n = 64 to estimate the mean of population with σ^2 = 2.56.

Sol :

Here n = 64

The probability = 0.90

$\sigma^2 = 2.56 \Rightarrow \sigma = \sqrt{2.56} = 1.6$

Confidence limit = 90%

$\therefore z_{\alpha/2} = 1.645$

Hence maximum error $E = z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} = 1.645 \times \frac{1.6}{\sqrt{64}} = 0.329$

3.5 T-DISTRIBUTION

Q15. Define Small sample tests and t-distribution.

Ans :

Small Sample test is one which consist of sample size less than (n<30) in small samples we can't assume the normal Distribution Approxmently it is denoted by "t" it is also called as "t" distribution in small sample test population size is not known in "t" test we assumed that the population from sample as been taken is normal.

Acceptance & Rejectence Criteria :

i) If $t_{cal} < t_{tab}$ Accept H_0

ii) If $t_{cal} > t_{tab}$ reject H_0

Here t_{cal} = calculated value

t_{tab} = tabulated value

If we take a very large number of small samples from a population and calculate the mean for each sample and then plot the frequency distribution of these means the resulting sampling distribution would be the **Student's t-distribution**.

The greatest contribution to the theory of small samples was made by **Sir William Gossett and R.A. Fisher**. Gossett published his discovery in 1905 under the pen name 'students' and its popularly known as t-test or students' t-distribution or students' distribution.

When the sample size is 30 or less and the population standard deviation is unknown, we can use the t-distribution.

The formula is, $t = \frac{(\bar{X} - \mu)}{S} \times \sqrt{n}$

Where S, or sample using n–1 as denominator then

$$t = \frac{(\bar{X} - \mu)}{S / \sqrt{n-1}} .$$

It should be noted that the only difference in the calculation of S in large and small samples is that whereas in the case of the former the sum of the squares of deviations of various items from the mean of $\Sigma(X - \bar{X})^2$ is divided by n–1. (the number of items) in case of small samples it is divided by n–1, which are the degrees of freedom.

The degrees of freedom in such problems is n–1 because one has the freedom to change only n–1 items as the last item has to be the difference between SX and sum of n–1 items. Thus, if there are 5 items in a sample with a total of 30. We can change only 4 items as we like. The last item will have to be 30 minus the sum of the remaining 4 items, whose values have been changed.

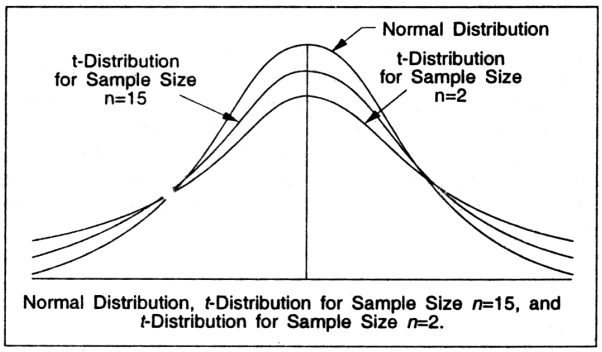
Q16. Explain the properties and applications of t-distribution.

Ans :

Properties

1. The variable t-distribution ranges from minus infinity to plus infinity.
2. The constant *c* is actually a function of *v* (pronounced as nu), so that for a particular value of *v*, the distribution of f(t) is completely specified. Thus f(t) is a family of functions, one for each value of *v*.
3. Like the standard normal distribution, the t-distribution is symmetrical and has a mean zero.
4. The variance of the t-distribution is greater than one, but approaches one as the number of degrees of freedom and, therefore, the sample size becomes large. Thus the variance of the t-distribution approaches the variance of the standard normal distribution as the sample size increases. It can be demonstrated that from an infinite number of degrees of freedom ($v = \infty$), the t-distribution and normal distribution are exactly equal. Hence there is a widely practised rule of thumb that samples of size n > 30 may be considered large and the standard normal distribution may appropriately be used as an approximation to t-distribution, where the latter is the theoretically correct functional form.

The following diagram compares one normal distribution with two t-distributions of different sample sizes :



Normal Distribution, *t*-Distribution for Sample Size *n*=15, and *t*-Distribution for Sample Size *n*=2.

The above diagram shows two important characteristics of t-distribution. First, a t-distribution is lower at the mean and higher at the tails than a normal distribution. Second, the t-distribution has proportionately greater area in its tails than the normal distribution. Interval widths from t-distributions are, therefore, wider than those based on the normal distribution.

The t- Table. The t-table given at the end is the probability integral of t-distribution. It gives, over a range of *v*. the probabilities of exceeding by chance value of t at different levels of significance. The t-distribution has a different value for each degree of freedom and when degrees of freedom are infinitely large, the t-distribution is equivalent to normal distribution and the probabilities shown in the normal distribution tables are applicable.

Applications of t-distribution

The following are some important applications of t-distribution,

(i) Test of hypothesis about the population mean.

(ii) Test of hypothesis about the difference between two means.

(iii) Test of hypothesis about the difference between two means with dependent samples.

(iv) Test of hypothesis about coefficient of correlation.

Q17. Explain the procedure for testing of single mean.

Ans :

In determining whether the mean of a sample drawn from a normal population deviates significantly from a stated value (the hypothetical value of the populations mean), when variance of the population is unknown we calculate the statistic :

$$t = \frac{(\bar{X} - \mu)\sqrt{n}}{S}$$

where $\bar{X}$ = the mean of the sample

μ = the actual or hypothetical mean of the population

n = the sample size

S = the standard deviation of the sample

$$S = \sqrt{\frac{\Sigma(X - \bar{X})^2}{n-1}}$$

Procedure of testing hypothesis :

Step 1 : Define null hypothesis (H_0)

Step 2 : Define Alternative hypothesis (H_1)

Step 3 : Deside level of significant

Step 4 : Test statistic

$$t = \frac{|\bar{x} - \mu|\sqrt{n}}{s}$$

$$s = \sqrt{\frac{\Sigma(x - \bar{x})^2}{n-1}}$$

$\bar{x}$ = Mean of the sample

μ = Mean of the population

s = Standard deviation of sample

n = Sample size

PROBLEMS

5. A manufacturer of certain electric Bulbs claims that is bulbs have mean life 25 months with standard deviation 5 months a random sample of 6 such bulbs areas follows.

Life of Bulbs	1	2	3	4	5	6
Months	24	26	30	20	20	18

Can you regard the procedure clamied to be valid at 1% Los ($t_{0.01}$ = 4.032).

Sol : **(Imp.)**

Given that

Mean life (μ) = 25 months

Random sample (n) = 6 bulbs (6 < 30)

Standard deviation (σ) = 5 months

Level of Significant = 1%

Step 1 : Null hypothesis : (H_0)

There is no significant difference between mean life of bulbs

$$H_0 = \mu_1 = \mu_2$$

Step 2 : Alternative hypothesis : (H_1)

There is significant difference between mean life of bulbs

$$H_1 = \mu_1 \neq \mu_2$$

Step 3 : Level of significant

Los = 1%

$t_{0.01}$ = 4.032

Step 4 : Test Statistics

"t" $\dfrac{|\bar{X} - \mu|\sqrt{n}}{S}$

$\bar{X}$ = Population mean

μ = Sample mean

n = Sample size

$$S = \sqrt{\frac{\Sigma(x - \bar{x})^2}{n-1}}$$

X	$X - \bar{X}$	$(X - \bar{X})^2$
24	24 – 23 = 1	1
26	26 – 23 = 3	9
30	30 – 23 = 7	49
20	20 – 23 = – 3	9
20	20 – 23 = – 3	9
18	18 – 23 = – 5	25
		$\Sigma(x - \bar{x})^2 = 102$

$$\bar{x} = \frac{\Sigma x}{N}$$

$$= \frac{138}{6}$$

$$\boxed{\bar{x} = 23}$$

$$s = \sqrt{\frac{\Sigma(x - \bar{x})^2}{n-1}}$$

$$s = \sqrt{\frac{102}{5}}$$

$$s = \sqrt{20.4}$$

$\boxed{s = 4.51}$

Substitute "S" value in "t"

$$\therefore t = \frac{|\bar{x} - \mu|}{s}\sqrt{n}$$

$$t = \frac{|23 - 25|\sqrt{6}}{4.51}$$

$$t_{cal} = \frac{|-2|\sqrt{6}}{4.51}$$

$$t_{cal} = \frac{(2)(2.44)}{4.51}$$

$$t_{cal} = \frac{4.88}{4.57}$$

$\boxed{t_{cal} = 1.08}$

$t_{cal} < t_{calse}$

$1.08 < 4.032$

$\therefore$ Accept the null hypothesis. (H_0)

$\therefore$ There is no significant different between mean life of bulbs.

6. A random sample of size 16 as 53 mean the sum of the squares of the deviation taken from the mean is 135. Can these sample regarded as taken from the population having 56 mean ? Obtain at 1% Los $t_{0.01} = 2.95$.

Sol :

Given that

Random sample size (n) = 16

Sample mean ($\bar{x}$) = 53

Mean sum of squares of the deviation = $\Sigma(x - \bar{x})^2$ = 135

Level of significance = 1% = $t_{0.01}$ = 135

Population mean (μ) = 56

Level of significance = 1% = $t_{0.01}$

Step (1) : The sample size is less than (30)

Hence it is considered as 't" test

Hypothesis :

Null Hypothesis : There is no significant difference b/w sample mean & population mean

$H_0 = \bar{x} = \mu$

Step 2 :

Alternative Hypothesis : There is significant difference b/w sample mean & population mean

$H_1 = \bar{x} \neq \mu$

Step 3 : Level of Significance

Loss = 1%

$t_{0.01}$ = 2.95

Degree of freedom (D.F) = n – 1

= 16 – 1

$\boxed{v = 15}$

Step 4 : Test statistic

$$t = \frac{|\bar{x} - \mu|\sqrt{n}}{s}$$

$$s = \sqrt{\frac{\Sigma(x - \bar{x})^2}{n-1}}$$

$$s = \sqrt{\frac{135}{15}}$$

$\boxed{s = 3}$

Substitute "S" value in "t"

$$t = \frac{|53 - 56|\sqrt{16}}{3}$$

$$t = {}^{(3)}\!/_{3}\,(4)$$

$\boxed{t = 4}$

Step 5 : Acceptance & Rejectence Criteria

$t_{cal} < t_{tab} \rightarrow$ Accept H_0

$t_{cal} > t_{tab} \rightarrow$ Reject H_0

$4 > 2.95 \rightarrow$ Reject H_0

$\therefore$ Accept Alternative Hypothesis (H_1)

$\therefore$ There is significant difference b/w population & sample means.

Q18. Explain the procedure for testing of two mean.

Ans :

Given two independent random samples of size n_1 and n_2 with means $\bar{X}_1$ and $\bar{X}_2$ and standard deviations S_1 and S_2 we may be interested in testing the hypothesis that the samples come from the same normal population. To carry out the test, we calculate the statistic as follows :

$$t = \frac{\bar{X}_1 - \bar{X}_2}{S} \times \sqrt{\frac{n_1 n_2}{n_1 + n_2}}$$

where $\bar{X}_1$ = mean of the first sample

$\bar{X}_2$ = mean of the second sample

n_1 = number of observations in the first sample

n_2 = number of observations in the second sample

S = combined standard deviation

The value of S is calculated by the following formula :

$$S = \sqrt{\frac{\Sigma\left(X_1 - \bar{X}_1\right)^2 + \Sigma(X_2 - \bar{X}_2)^2}{n_1 + n_2 - 2}}$$

PROBLEMS

7. Two different types of drugs 'A' and 'B' were tried on certain patients for increasing weight. 6 persons were given drug 'A" and 8 persons were given drug 'B'. The increase in pounds is given below,

Drug 'A'	**7**	**10**	**13**	**12**	**4**	**8**		
Drug 'B'	**12**	**8**	**3**	**18**	**16**	**9**	**8**	**3**

Do the drugs 'A' and 'B' differ significantly with regard to their effect in increase in weight?

Sol : **(Imp.)**

Let the weights (in kgs) of the patients treated with drugs A and B be denoted by suitable variables X and Y respectively.

We set up the null hypothesis, $H_0 : \mu_x = \mu_y$ i.e., there is no significant difference between the drugs A and B, regards their effect on increase in patients weight.

Alternative hypothesis, $H_1 : u_x \neq \mu_y$

Under H_0, the appropriate test statistic is,

$$t = \frac{\bar{x} - \bar{y}}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Degree of freedom (d.f) = $t_{n_1+n_2-2}$

Computation of Sample Means and Standard Deviations

Here, $n_1 = 6$, $\Sigma X = 54$, $\Sigma(X - \bar{x})^2 = 56$, $\bar{x} = \frac{\Sigma x}{n_1} = \frac{54}{6} = 9$

$n_2 = 8$, $\Sigma y = 77$, $\Sigma(y - \bar{y})^2 = 209.87$, $\bar{y} = \frac{\Sigma y}{n_2} = \frac{77}{8} = 9.625$

And, $S^2 = \frac{1}{n_1 + n_2 - 2}\left[\Sigma(x - \bar{x})^2 + \Sigma(y - \bar{y})^2\right]$

$= \frac{1}{6+8-2}$ [56 + 209 .87]

$= \frac{265.87}{12} = 22.16$

$\therefore S = \sqrt{22.16} = 4.71$

$$\therefore t = \frac{\bar{x} - \bar{y}}{S\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{9 - 9.625}{4.71\sqrt{\left(\frac{1}{6} + \frac{1}{8}\right)}} = \frac{-0.625}{4.71\sqrt{\frac{7}{24}}} = \frac{-0.625}{2.54} = -0.25$$

Degree of freedom (d.f) = $t_{n_1+n_2-2}$

$= t_{6+8-2} = t_{12}$

Hence, tabulated value of t for 12 d. f at 5% level of significance for the left tailed test is 1.782.

Thus, calculated t = – 0.25, which is less than tabulated value of t, (i.e.,1.782).

Therefore, null hypothesis H_0 cannot be rejected at 5% level of significance and we may conclude that the drugs A and B do not differ significantly as regards their effect on increases in patients weights at 5% level of significance.

Q19. Explain briefly about paired t-test.

Ans :

Paired observations arise in many practical situations where each homogeneous experimental unit receives both population conditions. As a result, each experimental unit has a pair of observations, one for each population.

For instance, to test the effectiveness of "drug" some 11 persons blood pressure is measured "before" and "after" the intake of certain drug. Here the individual person is the experimental unit and the two populations are blood pressure "before" and "after" the drug is given. Thus for each observation in one sample, there is a corresponding observation in the other sample pertaining to the same character. Hence the two samples are not independent.

Consider another example. Suppose a business concern is interested to know whether a particular media of promoting sales of a product is really effective or not. In this case we have to test whether the average sales before and after the sales promotion are equal.

If (x_1, y_1), (x_2,y_2),.....(x_n,y_n), be the pairs of sales data before and after the sales promotion in a business concern, we apply paired t - test to examine the significance of the difference of the two situations.

Let $d_t = x_i - y_i$ or $y_j - x_i$ for i = 1,2,3,....,n

Let the Null Hypothesis be $H_0 : \mu_1 = \mu_2$ (i.e., $\mu = 0$), there is no significant difference between the means in two situations.

Then the Alternative Hypothesis is $H_1: \mu_1, \neq \mu_2$

Assuming that H_0 is true, the test statistic for n paired observations (which are dependent) by taking the differences $d_x,d_2,....,d_n$ of the paired data.

$$t = \frac{\bar{d} - \mu}{S/\sqrt{n}} = \frac{\bar{d}}{S/\sqrt{n}} \quad (\because \mu = 0)$$

where $\bar{d} = \frac{1}{n}\Sigma d_1$ and $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(d_1 - \bar{d})^2$

are the mean and variance of the differences $d_y,d_2,.....,d_n$ respectively and μ is the mean of the population of differences.

The above statistic follows student's t-distribution with (n – 1) degrees of freedom.

PROBLEMS

8. Ten workers were given a training programme with a view to study their assembly time for a certain mechanism. The results of the time and motion studies before and after the training programme are given below

Workers	1	2	3	4	5	6	7	8	9	10
X_1	15	18	20	17	16	14	21	19	13	22
Y_1	14	16	21	10	15	18	17	16	14	20

X_1 = Time taken for assembling before training.

Y_1 = Time taken for assembling after training.

Test whether there is significant difference in assembly time before and after training.

Sol :

From the given paired data, we see that we are to use paired t-test.

Let μ be the mean of population of differences.

1. **Null Hypothesis H_0 :** $\mu_1 = \mu_2$ or m = 0 i.e., training is not useful.
2. **Alternative Hypothesis H_1 :** $\mu_1 \neq \mu_2$ i.e., training is useful in assembly time
3. **Level of significance :** $\alpha = 0.05$
4. **Computation :** Difference d_1's (before and after training) are

1, 2, –7, 1, –4, 4,3, –1,2

$\bar{d}$ = mean of differences of sample data = $\frac{\Sigma d}{n} = \frac{14}{10} = 1.4$

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(d_1 - \bar{d})^2$$

$$= \frac{1}{9}(1–1.4)^2 + (2 –1.4)^2 + (–1–1.4)^2 + (7–1.4)^2 + (1–1.4)^2 + (–4–1.4)^2 + (4–1.4)^2 + (3–1.4)^2 + (–1–1.43)^2 + (2 – 1.4)^2]$$

$$= \frac{1}{9}[0.16 + 0.36 + 5.76 + 31.36 + 0.16 + 29.16 + 6.76 + 2.56 + 5.76 + 0.36]$$

$$= \frac{82.4}{9} = 9.1555$$

$\therefore$ S = 3.026

5. **The test statistic is** $t = \frac{\bar{d} - \mu}{S/\sqrt{n}} = \frac{\bar{d}}{S/\sqrt{n}} = \frac{1.4}{3.026\sqrt{10}} = \frac{(1.4)(3.1623)}{3.026} = 1.46$

$\therefore$ Calculated |t| = 1.46

Tabulated $t_{0.05}$ with 10 – 1 = 9 degress of freedom is 1.833

Since calculated $t < t_{0.05,}$ we accept the Null Hypothesis H_0 and conclude that there is no significant difference in assembly times before and after training.

---

9. Score obtained in shooting competition by 10 soldiers before and after intensive training are given below :

Before	**67**	**24**	**57**	**55**	**63**	**54**	**56**	**68**	**33**	**43**
After	**70**	**38**	**58**	**58**	**56**	**57**	**68**	**75**	**42**	**38**

Test whether the intensive training is useful at 0.05 level of significance.

Sol :

Let us apply paired test

Let m be the mean of population of differences.

1. **Null Hypothesis H_0 :** $\mu_1 - \mu_2$ i..e, $\mu = 0$ there is no significant effect of the training.
2. **Alternative Hypothesis H_1 :** $\mu_1 - \mu_2$ intensive training is useful

3. Level of significance, $\alpha = 0.05$

4. Computation :

Differences d_i's (before and after training) are –3, –14, –1, –3,7, –13, –12, –7, –9,5

$$\bar{d} = \frac{1}{10}\sum_{i=1}^{10} d_1 = \frac{-50}{10} = -5$$

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(d_1 - \bar{d})^2 = \frac{1}{9}\sum_{i=1}^{10}(d_1 - \bar{d})^2$$

$$= \frac{1}{9}](2)^2 + (-9)^2 + (4)^2 + (2)^2 + (12)^2 + (-8)^2 + (-7)^2 + (-2)^2 + (-4)^2 + (10)^2]$$

$$= \frac{1}{9}[8+81+16+4+144+64+49+4+16+100]$$

$$= \frac{482}{9} = 53.5555$$

$\therefore S = 7.32$

5. The test statistic is $t = \frac{\bar{d}}{S/\sqrt{n}} = \frac{-5-0}{7.32/\sqrt{10}} = -2.16$

$\therefore |t| = 2.16$

Tabulated $t_{0.05}$ with 10 – 1 = 9 degree of freedom is 1.83

Since calculated t > t tabulated t, we reject the Null Hypothesis and conclude that the intensive training is useful.

UNIT IV

Testing Hypothesis - **one sample tests** - Hypothesis testing of mean when the population standard deviation is known, powers of hypothesis test, hypotheses testing of proportions, hypotheses testing of means when standard deviation is not known.

Testing Hypotheses - Two sample tests - Tests for difference between means - large sample, small sample, with dependent samples, testing for difference between proportions – Large sample.

4.1 Testing Hypothesis

Q1. Define Hypothesis. Explain the procedure for testing a Hypothesis.

Ans : **(Imp.)**

Statistical hypothesis is an assumption about the parameters of the population and sometimes it also concerns the type and nature of the distribution.

Example

(i) The average height of soldiers in the army is 165 cm.

(ii) A given drug cures 80% of the patients taking it.

(iii) A given machine has an effective life of 20 years.

All these hypotheses may be verified on the basis of certain sample tests. Procedures or tests which enable us to decide whether to accept or reject the hypothesis are called tests of hypothesis or tests of significance.

Procedure

Test of Hypothesis involves the following steps :

Step 1 : Statement (or assumption) of hypothesis

There are two types of hypothesis :

(i) Null Hypothesis

(ii) Alternative Hypothesis

(i) Null Hypothesis

For applying the tests of significance, we first set up a hypothesis a definite statement about the population parameter. Such a hypothesis is usually a hypothesis of no-difference, is called Null Hypothesis.

It is in the form $H_0 : \mu = \mu_0$

μ_0 is the value which is assumed or claimed for the population characteristic. It is the reference point against which the Alternative Hypothesis is set up, as explained in the next step.

Definition

A null hypothesis is the hypothesis which asserts that there is no significant difference between the statistic and the population parameter and whatever observed difference is there, is merely due to fluctuations in sampling from the same population. It is always denoted by H_0 . To test whether one procedure is better than another, we assume that there is no difference between the procedures. Similarly to test whether there is a relationship between two variates, we take H_0 that there is no relationship.

For example, in case of a single statistic, H_0 will be that the sample statistic does not differ significantly from the hypothetical parameter value and in the case of two statistics (H_0) will be that the sample statistics do not differ significantly.

(ii) Alternative Hypothesis

Any hypothesis which contradicts the Null Hypothesis is called an Alternative Hypothesis, usually denoted by H_1 The two hypothesis H_0 and H_1 are such that if one is true, the other is false and vice versa. For example, if we want to test the null hypothesis that the population has a specified mean μ_0 (say) i.e.,

$H_0 : \mu = \mu_0$, then the Alternative Hypothesis would be

(i) $H_1 : \mu \neq \mu_0$ (i.e., either $\mu > \mu_0$ or $\mu < \mu_0$)

or (ii) $H_1 : \mu > \mu_0$

or (iii) $H_1 : \mu < \mu_0$

The Alternative Hypothesis (i) is known as a two-tailed alternative and the Alternative Hypothesis in (ii) is known as right-tailed and in (iii) is known as left-tailed.

The setting of alternative hypothesis is very important to decide whether we have to use a single-tailed (right or left) or two-tailed test.

Alternate Hypothesis is in one of the following forms.

$H_1 : \mu \neq \mu_0$

or $H_1 : \mu > \mu_0$

or $H_1 : \mu < \mu_0$

Step 2 : Specification of the Level of Significance

The level of significance denoted by a is the confidence with which we rejects or accepts the Null hypothesis H_0 i.e., it is the maximum possible probability with which we are willing to risk an error in rejecting H_0 when it is true. The level of significance is generally specified before a test procedure so that the results obtained may not influence our decision. In practice, we take either 5% (i.e., 0.05) or 1% (i.e., 0.01) level of significance, although other levels such as 2%, 1/2% etc. may also be used. 5% Level of significance in a test procedure indicates that there are about 5 cases in 100 that we would reject the null hypothesis H_0 when it is true i.e., we are about 95% confident that we have made the right decision. Similarly, in 1% Level of significance, there is only 1 case in 100 that the null hypothesis H_0 is rejected when it is true i.e., we are about 99% confident that we have made the right decision. Level of significance is also known as the size of the test.

Step 3 : Identification of the Test Statistic

There are several tests of significance, viz., z,t,F etc. First we have to select the right test depending on the nature of the information given in the problem. Then we construct the test criterion and select the appropriate probability distribution.

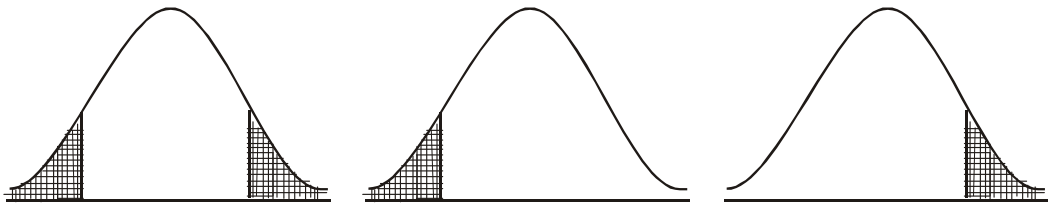
Step 4 : Critical Region

The critical region is formed based on following factors.

(a) Distribution of the Statistic i.e., whether the statistic follows the normal, 't', χ^2 or 'F' distribution (will be discussed later).

(b) **Form of Alternative Hypothesis**

If the form has $\neq$ sign, the critical region is devided equally in the left and right tailes, sides of the distribution.



Two sided Left sided Right sided

If the form of alternative hypothesis has < sign, the entire critical region is taken in the left tail of the distribution.

If the form of alternative hypothesis has > sign, the entire critical region is taken on the right side of the distribution.

Step 5 : Making Decision

We compute the value of the appropriate statistic and ascertain whether the computed value falls in acceptance or rejection region depending on the specified Level of significance. In finding the acceptance or rejection region we have to use critical values given in Statistical Tables. By comparing the computed value and the critical value decision is taken for accepting or rejecting H_0. If the computed value < critical value, we accept H_0, otherwise we reject H_0.

Q2. Explain the types of errors in sampling.

Ans :

The main objective in sampling theory is to draw valid inferences about the population parameters on the basis of the sample results. In practice we decide to accept or to reject the lot after examining a sample from it. As such we have two types of errors.

(i) Type I error : Reject H_0 when it is true.

If the Null Hypothesis H_0 is true but it rejected by test procedure, then the error made is called Type I error or α error.

(ii) Type II error : Accept H_0 when it is wrong i.e., accept H_0 when H_1 is true.

If the Null Hypothesis is false but it is accepted by test, then error committed is called Type II error or β error.

If we write

P (Reject H_0 when it is true)

= P(Type I error) = α

and P (Accept H_0 when it is wrong)

= P (Type II error) = β

then α and β are called sizes of Type I and Type II errors respectively

i.e., α = P (Rejecting a good lot)

β = P (Accepting a bad lot)

The sizes of Type I and Type II errors are also know as producer's risk and consumer's risk respectively.

4.1.1 One Sample Tests

Q3. Explain the concept One Sample Test of Hypothesis

Ans :

The one sample z test isn't used very often because we rarely know the actual population standard deviation . However, it's a good idea to understand how it works as it's one of the simplest tests you can perform in hypothesis testing. In English class you got to learn the basics (like grammar and spelling) before you could write a story; think of one sample z tests as the foundation for understanding more complex hypothesis testing.

$$Z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$$

PROBLEMS

1. A principal at a certain school claims that the students in his school are above average intelligence. A random sample of thirty students IQ scores have a mean score of 112. Is there sufficient evidence to support the principal's claim? The mean population IQ is 100 with a standard deviation of 15.

Sol :

Step 1

State the Null hypothesis. The accepted fact is that the population mean is 100, so: H_0: y = 100.

Step 2

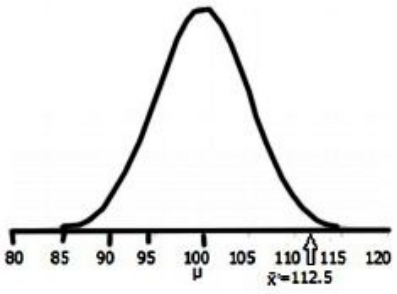
State the Alternate Hypothesis. The claim is that the students have above average IQ scores, so:

$$H_1 : \mu > 100$$

The fact that we are looking for scores "greater than" a certain point means that this is a one-tailed test.

Step 3

Draw a picture to help you visualize the problem.



Step 4

State the alpha level. If you aren't given an alpha level, use 5% (0.05).

Step 5

Find the rejection region area (given by your alpha level above) from the z-table. An area of .05 is equal to a z-score of 1.645.

Step 6

Find the test statistic using this formula:

$$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$$

Step 7

If Step 6 is greater than Step 5, reject the null hypothesis. If it's less than Step 5, you cannot reject the null hypothesis. In this case, it is greater (4.56 > 1.645), so you can reject the null.

---

2. Blood glucose levels for obese patients have a mean of 100 with a standard deviation of 15. A researcher thinks that a diet high in raw cornstarch will have a positive or negative effect on blood glucose levels. A sample of 30 patients who have tried the raw cornstarch diet have a mean glucose level of 140. Test the hypothesis that the raw cornstarch had an effect.

Sol :

Step 1

State the null hypothesis: $H_0: \mu = 100$

Step 2

State the alternate hypothesis: $H_1: \neq 100$

Step 3

State your alpha level. We'll use 0.05 for this example. As this is a two-tailed test, split the alpha into two.

$0.05/2 = 0.025$

Step 4

Find the z-score associated with your alpha level. You're looking for the area in one tail only. A z-score for 0.75(1– 0.025 = 0.975) is 1.96. As this is a two-tailed test, you would also be considering the left tail (z = 1.96)

Step 5

Find the test statistic using this formula:

$$Z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$$

$$z = (140 - 100)/(15 / \sqrt{30})$$

$$= 14.60.$$

Step 6

If Step 5 is less than -1.96 or greater than 1.96 (Step 3), reject the null hypothesis. In this case, it is greater, so you can reject the null.

4.1.2 Hypothesis Testing of Mean when the Population Standard Deviation is known

Q4. How Hypothesis testing of mean can be done when the population standard deviation is them known with an examples.

Ans : **(Imp.)**

We can also use the standard normal distribution, or z-scores, to test a mean hypothesis when the population standard deviation is known. The next two examples, though they have a smaller sample size, have a known population standard deviation.

Example 1:

A sample of size 50 is taken from a normal distribution, with a known population standard deviation of 26. The sample mean is 167.02. Use the 0.05 significance level to test the claim that the population mean is greater than 170.

We always put equality in the null hypothesis, so our claim will be in the alternative hypothesis.

$H0 : \mu = 170$

$HA : \mu > 170$

The test statistic is:

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} = \frac{167.02 - 170}{26 / \sqrt{50}} \approx \frac{2.98}{3.67} = 0.81$$

Now we will find the probability of observing a test statistic at least this extreme when assuming the null hypothesis. Since our alternative hypothesis is that the mean is greater, we want to find the probability of z scores that are greater than our test statistics. The p-value we are looking for is:

$$\text{p-value} = P(z > 0.811)$$

$$= 1 - P(z < 0.811)$$

$$= 1 - 0.791$$

$$= 0.209 > 0.05$$

The probability of observing a test statistic at least as big as the $z = 0.81$ is 0.209. Since this is greater than our significance level, 0.05, we fail to reject the null hypothesis. This means that the data does not support the claim that the mean is greater than 170.

Example 2:

A sample of size 20 is taken from a normal distribution, with a known population standard deviation of 0.01. The sample mean is 0.194. Use the 0.01 significance level to test the claim that the population mean is equal to 0.22.

We always put equality in the null hypothesis, so our claim will be in the null hypothesis. There is no reason to do a left or right tailed test, so we will do a two tailed test:

$H0 : \mu = 0.22$

$HA : \mu \neq 0.22$

The test statistic is:

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} = \frac{0.194 - 0.22}{0.04 / \sqrt{20}}$$

$$\approx \frac{-0.026}{0.008944} = -2.91$$

Now we will find the probability of observing a test statistic at least this extreme when assuming the null hypothesis. Since our alternative hypothesis is that the mean is not equal to 0.22, we need to find the probability of being less than -2.91, and we also need to find the probability of being greater than positive 2.91. However, since the normal distribution is symmetric, these probabilities will be the same, so we can find one and multiply it by 2:

$$\text{p-value} = 2 \oplus P(z < -2.91)$$

$$= 2 \oplus 0.0018 = 0.0036 / > 0.01$$

The probability of observing a test statistic at least as extreme as $z = -2.91$ is 0.0036. Since this is less than our significance level, 0.01, we reject the null hypothesis. This means that the data does not support the claim that the mean is equal to 0.22.

A sample of size 36 is taken from a normal distribution, with a known population standard deviation of 57. The sample mean is 988.93. Use the 0.05 significance level to test the claim that the population mean is less than 1000.

We always put equality in the null hypothesis, so our claim will be in the alternative hypothesis:

$H0: \mu = 1000$

$HA : \mu < 1000$

The test statistic is:

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} = \frac{988.93 - 1000}{57 / \sqrt{36}} \approx \frac{-11.07}{9.5}$$

$$= -1.17$$

Now we will find the probability of observing a test statistic at least this extreme when assuming the null hypothesis. Since our alternative hypothesis is that the mean is less than 1000, we need to find the probability of z scores less than -1.17:

p-value = P(z < –1.17) = 0.1210 > 0.05

The probability of observing a test statistic at least as extreme as $z = -1.17$ is 0.1210. Since this is greater than our significance level, 0.05, we fail to reject the null hypothesis. This means that the data does not support the claim that the mean is less than 1000.

4.1.3 Powers of Hypothesis Test

Q5. Elaborate the various Powers of hypothesis test.

Ans :

Meaning of β and 1 - β

Ideally, α and β (the probabilities of Type I and Type II errors) should both be small. Recall that a Type I error occurs when we reject a null hypothesis that is true, and that a (the significance level of the test) is the probability of making a Type I error. In other words, once we decide on the significance level, there is nothing else we can do about α. A Type II error occurs when we accept a null hypothesis that is false; the probability of a Type II error is β.

Suppose the null hypothesis is false. Then managers would like the hypothesis test to reject it all the time. Unfortunately, hypothesis tests cannot be foolproof; sometimes when the null hypothesis is false, a test does not reject it, and thus a Type II error is made. When the null hypothesis is false, μ. (the true population mean) does not equal μ_{H_0} (the hypothesized population mean); instead, μ equals some other value. For each possible value of μ for which the alternative hypothesis is true, there is a different probability (β) of incorrectly accepting the null hypothesis. Of course, we would like this β (the probability of accepting a null hypothesis when it is false) to be as small as possible, or, equivalently, we would like 1–β (the probability of rejecting a null hypothesis when it is false) to be as large as possible.

4.1.4 Hypothesis Testing of Proportions

Q6. Write about hypothesis testing - single proportion.

Ans :

The population parameter of interest is Population Proportion P. If the sample size is large than sample proportion '*p*' will be approximately normally distributed. Then

$$Z = \frac{p - P}{\sigma_p} \sim N(0,1)$$

If *x* is the number of individuals possessing the given attribute (which is termed a success) in a random sample of size *n* from an infinite population, then *p* = observed sample proportion of success = $\frac{X}{N}$. It has been stated that $E(p) = P$.

$$\text{S.E.}(p) = \sqrt{\frac{PQ}{n}}$$

where *P* is the population proportion of success. Hence for large samples, the standard normal variate corresponding to statistic '*p*' is

$$Z = \frac{p - E(p)}{S.E.(p)} = \frac{p - P}{\sqrt{\frac{PQ}{n}}} \sim N(0,1)$$

Test the population proportion having a specified value P_0 (say), i.e., $p = P_0$.

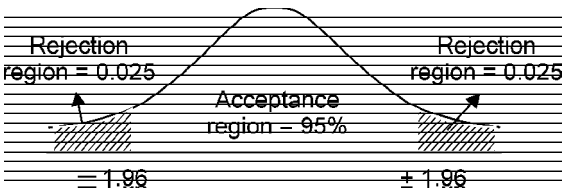
PROBLEMS

3. In 400 throws of six-faced dice, odd points appeared 260 times. Would you say that the dice is fair at 5% of significance?

Sol :

Let us take the hypothesis that the dice is unbiased, i.e.,

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



n = 400, x = 260, $np = 400\left(\frac{1}{2}\right) = 200$

The $H_0 : p = P$

$H_1 : p \neq P$ and a = 0.05

Applying the formula

$$Z = \frac{x - np}{\sqrt{npq}} = \frac{260 - 200}{\sqrt{400 \times \frac{1}{2} \times \frac{1}{2}}} = \frac{60}{\sqrt{100}} = 6$$

Conclusion

Since the computed value of Z = 6 is greater then the table value Z_a = ± 1.96 at 5% level of significance, the hypothesis is rejected. Hence the dice does not seem to be fair, i.e., it is biased.

4. A dice was thrown 9000 times and either 3 or 4 found to have been occurred 3220 times. Is this consistent with the hypothesis that the dice is unbiased ? Use a = 1%.

Sol :

The null hypothesis is that the dice was unbiased, i.e.,

$H_0 : p = P_0$

where P = Population proportion of success against the alternative hypothesis that the dice is not unbiased, i.e.,

$H_1 = p \neq P_0\left(1 = \frac{1}{3}\right)$ (Two – tailed test)

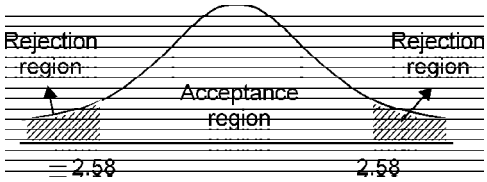
$P_0 - P_r$ (getting a 3 or 4 in throw of an unbiased dice)

$$P_0 = \frac{2}{6} = \frac{1}{3} = 0.3333$$

$$\Rightarrow \quad Q_0 = 1 - P_0 = 1 - \frac{1}{3}$$

= 1 – 0.3333

= 0.6667



We are given that n = 9000 and X = 3220, then the sample proportion p will be

$$p = \frac{X}{n} \text{ or}$$

$$p = \frac{3220}{9000} = 0.3578$$

If a = 0.01, then the test statistic Z will be

$$Z = \frac{p - P_0}{\sqrt{\frac{P_0 Q_0}{n}}} \sim N(0,1)$$

(under Null Hypothesis)

$$Z = \frac{0.3578 - 0.3333}{\sqrt{\frac{0.3333 \times 0.6667}{9000}}}$$

$$= \frac{0.0245}{\sqrt{0.0000246}} = \frac{0.0245}{0.00496} = 4.94$$

Since the calculated value of $|Z|$ = 4.94 which is greater than the tabulated value of Z_α = 2.58, we reject H_0. Hence, we conclude that the dice is certainly biased.

5. In a city A, out of 600 men, 325 men were found to be smokers. Does this information support the statement that 'Majority of men is this city are smokers'? Draw your conclusion using a = 1% and a = 5% separately.

Sol :

We are given that n = 600, X = 325 Sample proportion of smokers is

$$p = \frac{X}{n} = \frac{325}{600} = 0.5417$$

The null hypothesis is that the number of smokers and non-smokers are equal in the city so that P_0 = Population proportion of smokers in the city = $\frac{1}{2} = 0.5$

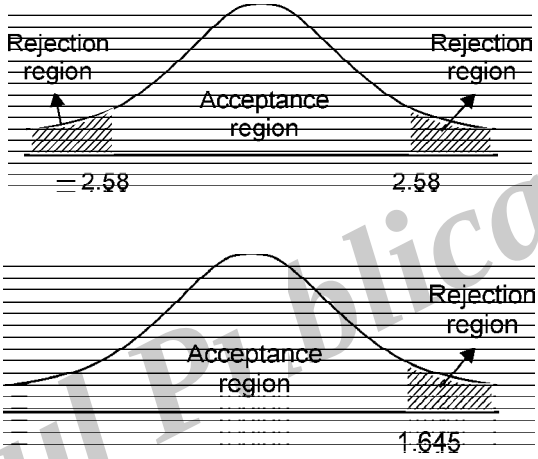
$$Q_0 = 1 - P_0 = 1 - \frac{1}{2} = 0.5$$

i.e., $H_0 : p = P_0 = \frac{1}{2} = 0.5$

Against the alternative hypothesis is that the majority of men in the city are smokers i.e.,

$p > P_0$ (= 0.5) (Right tailed) and

x = 0.01 and a = 0.05 then the test statistic will be



$$Z = \frac{p - P_0}{\sqrt{\frac{P_0 Q_0}{n}}} \sim N(0,1) \text{ (under } H_0)$$

$$Z = \frac{0.5417 - 0.5}{\sqrt{\frac{0.5 \times 0.5}{600}}}$$

$$= \frac{0.0417}{\sqrt{0.0004165}} = \frac{0.0417}{0.0204} = 2.04$$

Since the calculated value of Z = 2.04 is less than the tabulated value of Z_α = 2.33. We may accept H_0 at 1% level of significance. Hence we may conclude that the number of smokers and non-smokers are equal in the city.

b) Since the calculated value of Z = 2.04, is greater than the tabulated value of Z_α = 1.645 at 5% level of significance. We reject H_0. Hence, we conclude that the majority of men in the city are smokers.

Q7. Write about Hypothesis testing - Difference of proportions.

Ans : **(Imp.)**

Case - I (Population Proportions are Unequal)

Suppose that we want to compare two large populations with respect tot he prevalence of a certain attributes among their members. We have taken two independent random samples of sizes n_1 and n_2 respectively from these two populations. Let x_1 and x_2 be the observed number of success in the samples respectively. Then

P_1 = observed

proportion of success in the first sample = $\frac{x_1}{n_1}$

P_2 = observed

proportion of success in the second sample = $\frac{x_2}{n_2}$

If the population proportions are P_1 and P_2 then $E(p_1)$, $E(p_2) = P_2$ and variance of p_1 is $V(p_1) = \frac{P_1Q_1}{n_1}$ and variance of p_2 is $V(p_2) = \frac{P_2Q_2}{n_2}$

Let 't' be

Then the statistic for $(P_1 - P_2)$

$$E(t) = E(p_1 - p_2) = E(p_1) - E(p_2) = P_1 - P_2$$

$$V(t) = V(p_1 - p_2) = V(p_1) + V(p_2)$$

S.E.(t) = $S.E(P_1 - P_2) = \sqrt{V(P_1 - P_2)} = \sqrt{\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}}$

Hence the statistic for $(P_1 - P_2)$ is given by

$$Z = \frac{(p_1 - p_2) - E(p_1 - p_2)}{S.E.(p_1 - p_2)} = \frac{(p_1 - p_2) - (P_1 - P_2)}{\sqrt{\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}}} \sim N(0,1) \text{ (under } H_0\text{)}.$$

Case-II (Population Proportions are Equal)

If we want to test whether two independent random samples have come from same population, i.e., $H_0 : P_1 = P_2 = P$ (say)

$$Z = \frac{p_1 - p_2}{\sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0,1) \text{ (under } H_0\text{)}$$

where $\hat{p} = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$

PROBLEMS

6. In a simple random ample of 600 men taken from a big city 450 are found to be the users of a product 'A'. In another simple random sample of 900 men taken from another city 450 are found to be the users of a product 'A'. Do the data indicate that there is a significant difference in the habit of usage of the product 'A' in the two cities?

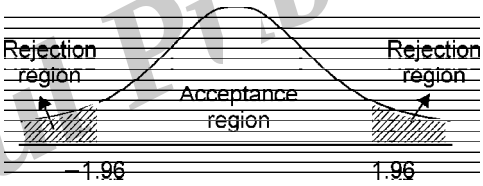
Sol :

We are given that n_1 = 600; X_1 = 450 ; n_2 = 900 and X_2 = 450

$$\Rightarrow \quad p_1 = \frac{X_1}{n_1} = \frac{450}{600} = 0.75$$

$$\Rightarrow \quad p_2 = \frac{X_2}{n_2} = \frac{450}{900} = 0.5$$

The null hypothesis is that the proportion of users of the product A in both the cities are equal.



i.e., H0 : $P_1 = P_2$

against $H_1 : P_1 \neq P_2$ (two tailed test).

Let a = 0.05, then the test statistic will be

$$Z = \frac{(p_1 - p_2) - (P_1 - P_2)}{\sqrt{\frac{P_1 Q_1}{n_1} + \frac{P_2 Q_2}{n_2}}} \sim N(0,1)$$

$$Z = \frac{(0.75 - 0.5) - (0)}{\sqrt{\frac{0.75 \times 0.25}{600} + \frac{0.5 \times 0.5}{900}}} = \frac{0.25}{\sqrt{0.00059}} = \frac{0.25}{\sqrt{0.02429}}$$

= 10.29

Since the calculated value of Z = '|10.29|' which is much greater than the tabulated value of Z_a = 1.96, we rejected H_0. Hence, we conclude that there is a significant difference between proportion of users of the product A in both the cities.

7. A manufacturer of pet animal foods was wondering whether cat owners and dog owners reacted differently to premium pet foods. They commissioned a consumer survey that yielded the following data.

Pet	Number of owners Surveyed	Number of owners using Premium food
Cat	280	152
Dog	190	81

is it reasonable to conclude at a = 0.05, the cat owners are more likely to feed their pets. Premium food than dog owners?

Sol : **(Imp.)**

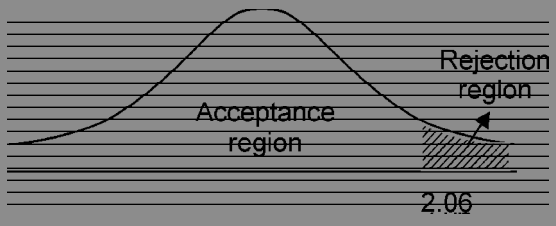
We are given following information :

$n_1 = 280; \quad X_1 = 152; n_2 = 190; \quad X_2 = 81$, then

$$p_1 = \frac{X_1}{n_1} = \frac{152}{280} = 0.549,$$

$$p_2 = \frac{X_2}{n_2} = \frac{81}{190} = 0.426$$

The null hypothesis is that there is no significance difference between cat owners and dog owners in feeding premium food to their pets, i.e., $H_0 : P_1 = P_2$, against the alternative hypothesis that cat owners are more likely to feed their pets that dog owners in feeding premium food their pets, i.e., $H_1 : P_1 > P_1$ (right tailed test) with a = 0.02



Then the test statistic will be

$$Z = \frac{p_1 - p_2}{\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}} \sim N(0,1) \text{ (under } H_0).$$

Since P_1 and P_2 are unknown, their unbiased, estimates, namely, p_1 and p_2 are being used.

$$Z = \frac{0.54300.426}{\sqrt{\frac{0.543 \times 0.457}{280} + \frac{0.426 \times 0.574}{190}}}$$

$$= \frac{0.117}{\sqrt{0.00217}} = \frac{0.117}{0.0466} = 2.51$$

$$= \frac{0.117}{0.0466}$$

$$= 2.51$$

Since the calculated value of $Z = 2.51$ which is greater than the tabulated value of $Z_\alpha = 2.06$ at 2% level of significance, we reject H_0. Hence, we conclude that the cat owners are more likely than dog owners to feed their pets with premium food.

4.1.5 Hypothesis Testing of means when Standard Deviation is not known.

Q8. Write about Hypothesis testing - single mean.

Ans :

Let $X_1, X_2.....X_n$ be a random sample of size n from a population with mean m and variance σ^2. Assume that σ^2 is known and μ is unknown. In developing the test of significance for a single mean, we are interested, to test if the mean of the population has a specified value μ_0(say), i.e., $\mu = \mu_0$.

One Tailed and Two Tailed Tests

In any test, the critical region is represented by a portion of the area under the probability curve of the sampling distribution of the test statistic. Basically there are three kinds of problems associated in testing of hypothesis. They are :

i) Two-tailed test

ii) Right-tailed test and

iii) Left-tailed test

i) Two-tailed test is a test where the alternative hypothesis leads to have the critical region in two-tails of the distribution :

$H_0 : m_1 = m_2$ (or $m = m_0$)

against $H_1 : \mu_1 \neq \mu_2$ (or $\mu \neq \mu_0$)

is known as two-tailed test. The diagram which presents the critical region into two tails is shown in fig. below.

One-Tailed test is a test where the alternative hypothesis leads to have the critical region in right side or left side of the distribution. This is also usually called as one (single) tailed test.

ii) Right-tailed test : $H_0 : m_1 = m_2$ (or $m = m_0$)

against $H_1 : m_1 > m_2$ (or $m = m_0$) Right-tailed test

The diagram which presents the critical region in Right tail is shown in figure below.

iii) Left-tailed test : $H_0 : m_1 = m_2$ (or $m = m_0$)

Against $H_1 : m_1 < m_2$ (or $m < m_0$) Left-tailed test. The diagram which presents the critical region in Left is shown in fig.below. The following diagrams would make it more char two-tailed test.

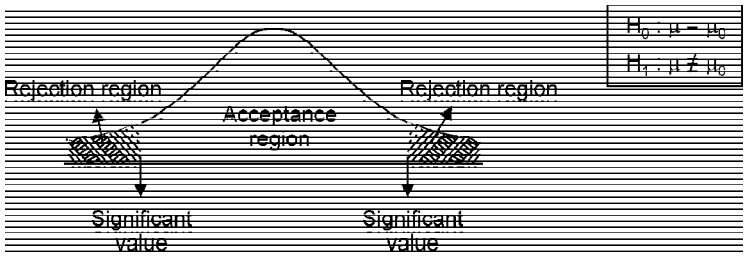


Fig.: Two-Tailed Test

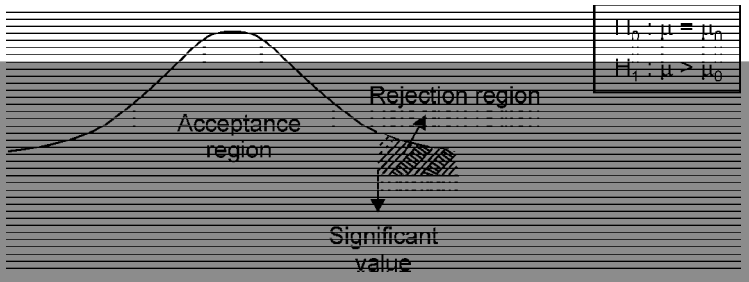


Fig.: Right-Tailed Test

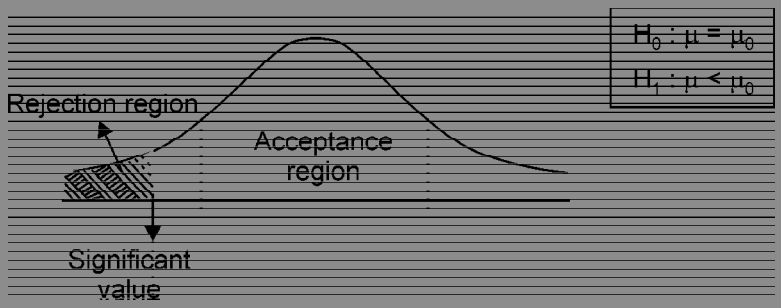


Fig.: Left-Tailed Test

PROBLEMS

8. The mean lifetime of a sample of 100 light tubes produced by a company is found to be 1,560 hours with a population standard deviation of 90 hours. Test the hypothesis that the mean lifetime of the tubes produced by the company is 1580 hours.

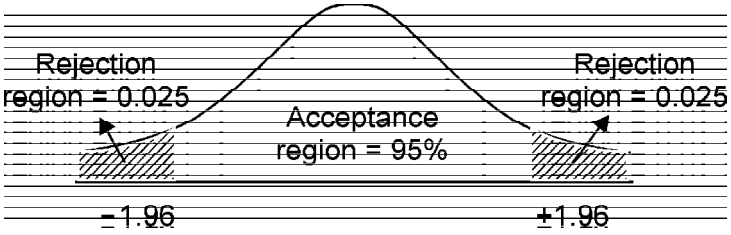
Sol :

The null hypothesis is that there is no significant difference between the sample mean and hypothetical population mean. The population mean has a specified value.

Null Hypothesis (H_0) : $m = m_0$

Alternative Hypothesis (H_1) : $m \neq m_0$

Let $a = 0.05$



Then the test statistics

$$Z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \sim N(0.1)$$

$$Z = \frac{1560 - 1580}{90/\sqrt{100}} = -\frac{20}{90/10} = -\frac{20}{9} = -2.22$$

Since it is two-tailed test, the critical value of $Z_a = \pm 1.96$ for a two-tailed test at a = 5% level of significance.

Conclusion

Since the computed value of $Z = -2.22$ in the rejection region, we reject the null hypothesis. Hence the mean lifetime of the tubes produced by the company may not be 1,580 hours.

9. The mean weight of 200 male students in a college is 62 kgs with a standard deviation of 4 kgs. Test the hypothesis that the mean weight in the population is greater than 58 kgs. Use a = 1%.

Sol :

The null hypothesis is that the population mean has a specified value, i.e.,

H_0 : m = 58 kgs (= m_0)

against H_1 : m > 58 kgs

a = 0.01

the test statistic

$$Z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$$

$$Z = \frac{60 - 58}{4/\sqrt{200}} = \frac{2}{4/14.14} = \frac{28.28}{4} = 7.07$$

The critical value of $Z_a = 2.33$ for a right-tailed at 1% level of significance.

Conclusion

Since computed value of $Z = 7.07$ falls in the rejection region, we reject the null hypothesis. Hence, the mean weight of the students is greater than 58 kgs.

4.2 Two Sample Tests

Q9. Write about Hypothesis testing of two means.

Ans : **(Imp.)**

Case - I (Population Variances are Unequal)

Let us consider two independent random samples of size n_1 and n_2 from the two populations with means m_1 and m_2 and variances σ_1^2 and σ_2^2 respectively. Let $\bar{X}_1$ and $\bar{X}_2$ be the corresponding sample means. Since $\bar{X} \sim N(\mu_1, \sigma_1^2/n_1)$ and $\bar{X}_2 \sim N(\mu_2, \sigma_2^2/n_2)$ we have $(\bar{X}_1 - \bar{X}_2)$ the difference of two independent normal variates is also a normal variate with mean $(m_1 - m_2)$ and variance $\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$

Therefore the standard normal variate

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - E(\bar{X}_1 - \bar{X}_2)}{S.E.(\bar{X}_1 - \bar{X}_2)} \text{ or}$$

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - 1(\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

Hence the problem is: To test the equality of the two population means, i.e., to test whether $m_1 = m_2$

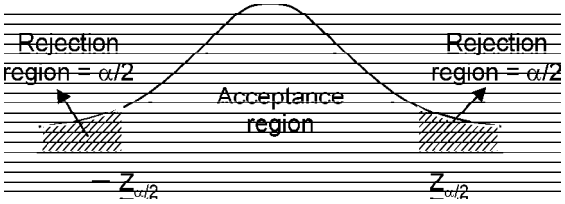
Type (a) – Two Tailed Test

Null hypothesis (H0) : $m_1 = m_2$

Alternative Hypothesis (H_1) : $\mu_1 \neq \mu_2$

(Two – tailed test)

Let a be the level of significance



Then the test statistic

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

(under H_0)

If $\sigma_{(\bar{X}_1 - \bar{X}_2)}$ is known then

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim N(0,1)$$

(for large samples)

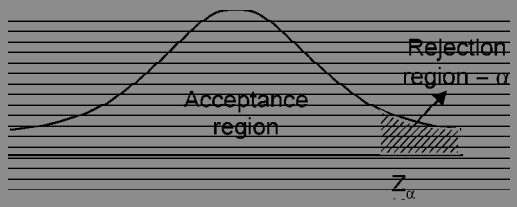
By comparing the calculated values of *Z* with the tabulated value of Z_a = 2.58, 1.96 or 1.645 (for two tailed) we can reject or retain the null hypothesis at 1%, 5% or 10% level of significance respectively.

Type (b) – Right tailed Test

If the hypothesis involves are right tailed test, i.e.,

$H_0 : \mu_1 = \mu_2$ and

$H_1 : \mu_1 > \mu_2$



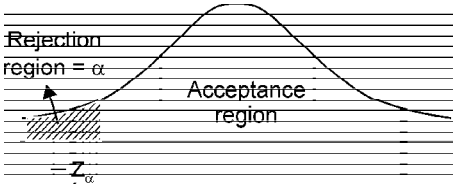
By comparing the calculated value of *Z* with the tabulated value of Z_a = 2.33, 1.645 and 1.28 (for right-tailed test) we can reject or retain the null hypothesis at 1%, 5% or 10% level of significance respectively.

Type (c) – Left Tailed Test

If the hypothesis involves a left-tailed test, i.e.,

$H_0 : \mu_1 = \mu_2$ and

$H_1 : \mu_1 < \mu_2$



By comparing the calculated value of Z with the tabulated value of Z_a = –2.33, –1.645 or –1.28 (for left-tailed test) we can reject or retain the null hypothesis at 1%, 5% or 10% level of significance respectively.

PROBLEMS

10. Two types of new cars produced in India are tested for petrol mileage. One sample consisting of 36 cars averaged 13 kmpl (kilometers per litre). While the other sample consisting of 72 cars averaged 11.5kmpl. with population variances as $\sigma_1^2 = 1.5$ and $\sigma_2^2 = 2.0$ respectively. Test whether there is any significant difference in the petrol consumption of these two types of cars. Use a = 0.01.

Sol :

n_1 = 36 $\bar{X}_1 = 13$ $\sigma_1^2 = 1.5$

n_2 = 72 $\bar{X}_2 = 11.5$ $\sigma_2^2 = 2.0$

We are given the following information

Null hypothesis

There is no significant difference in the petrol consumption of the two types of cars, i.e.,

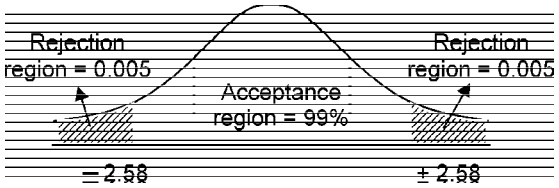
$H_0 : \mu_1 = \mu_2$; $H_1 : \mu_1 \neq \mu_2$

and a = 0.01 (given)

The test statistic

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1) \text{ (under } H_0)$$

$$Z = \frac{13 - 11.5}{\sqrt{\frac{1.5}{36} + \frac{2.0}{72}}} = \frac{1.5}{0.262} = 5.72$$



Conclusion

Since the calculated value of Z = 5.68 which is greater the tabulated value of Z_a = 2.58 at 1% level of significance, the null hypothesis is rejected. Hence there is a significant difference in the petrol consumption of the two types of cars.

11. Two brands of bulbs are quoted at the same price. A buyer tested a random sample of 60 bulbs of Brand A which gave a mean life time of 86 hours with a standard deviation of 6 hours. Another sample of 75 bulbs of Brand B which gave a mean life time of 82 hours with a standard deviation of 9 hours. Test whether the two brands of bulbs are same with regard to their average life? Use *a* = 0.10.

Sol :

We are given the following information.

n_1 = 60 $\bar{X}_1 = 86$ hours s_1 = 6 hours (sample standard deviation)

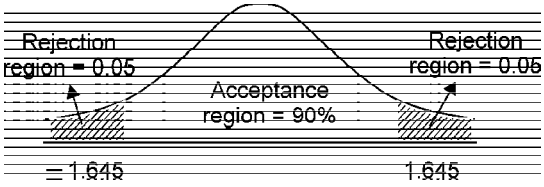
n_2 = 75, $\bar{X}_2 = 82$ hours s_2 = 9 hours (sample standard deviation)

Null hypothesis

The two brands are similar with regard to their average lifetimes. i.e $H_0 : m_1 = m_2$ against the alternative hypothesis that the samples have come from different populations.

$H_0 : m_1 = m_2$

$H_1 : \mu_1 \neq \mu_2$ and a = 0.10



Test statistic

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim N(0,1)$$

by using Central Limit Theorem

for large sample theory

$$Z = \frac{86-82}{\sqrt{\frac{36}{60} + \frac{81}{75}}} = \frac{4}{1.296} = 3.09$$

Conclusion

Since the calculated value of Z = 3.09 which is greater than the tabulated value of Z_a = ± 1.645 at 10% level of significance, we reject H_0. Hence we conclude that the two brands of bulbs are not same with regard to their average lifetimes.

12. The Financial Accounting Board (FAB) was considering a proposal to require companies to report the potential effect of employees' stock options on earnings per share (EPS). A random sample of 41 high-technology firms revealed that the new proposal would reduce EPS by an average of 13.8 per cent, with a standard deviation of 18.9 per cent.

A random sample of 35 producers of consumer goods showed that the proposal would reduce EPS by 9.1 per cent on average with a standard deviation of 8.7 percent. On the basis of these samples, can it be reasonable to conclude at a = 0.05 that the FAB proposal will causes a greater reduction in EPS for high-technology firms than for producers of consumer goods?

Sol :

We are given the following information :

Sample 1 (HT firms) :

n_1 = 41; $\bar{X}_1 = 13.8$ and s_1 = 18.9

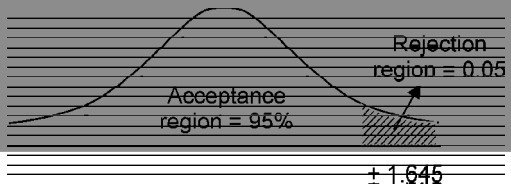
Sample 2 (CG producers) :

n_2 = 35 ; $\bar{X}_2 = 9.7$ and s_2 = 8.7

The null hypothesis is that both the sample have same population mean, i.e., $H_0 : \mu_1 = \mu_2$; against the alternative hypothesis that the first population mean is greater than the second population mean, i.e., $H_1 : \mu_1 > \mu_2$ (Right-tailed test) and a = 0.05, then the test statistic will be

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim N(0,1) \text{ under } H_0$$

(For large samples, since σ_1^2 and σ_2^2 are unknown and their estimates are used).



$$Z = \frac{13.8 - 9.1}{\sqrt{\frac{(18.9)^2}{41} + \frac{(8.7)^2}{35}}} = \frac{4.7}{3.29} = 1.43$$

Since the calculated value of Z = 1.43 is less than the tabulated value of Z_a = ± 1.645 at 5% level of significance, we may accept H_0. Hence, we conclude that there is no difference between FAB proposal in EPS for high-tech firms and consumer goods producers.

4.3 TESTS FOR DIFFERENCE BETWEEN MEANS - LARGE SAMPLE

Q10. Explain Two tailed test for difference between the means of two samples.

Ans :

If two independent random samples with m and n_2 numbers (both sample sizes are greater than 30) respectively are drawn from the same population of standard deviation 01 the standard error of the difference between the sample means is given by the formula :

S.E. of the difference between sample means

$$= \sqrt{\sigma^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$$

If σ is unknown, sample standard deviation for combined samples must be substituted.

If two random samples with $\bar{X}_1, \sigma_1, n_1$, 01, ni and $\bar{X}_2, \sigma_2, n_2$ respectively are drawn from different populations, then the S.E. of the difference between the mean is given by the formula :

$$= \sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$$

and where a_1 and a_2 are unknown.

S.E. of the difference between means

$$= \sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}$$

where S_1 and S_2 represent standard deviations of the two samples. The null hypothesis to be tested is that there is no significant difference in means of the two samples, i.e.,

$H_0 : \mu_1 = \mu_2 \leftarrow$ null hypothesis, there is no difference ,

$H_1 : \mu_1 \neq \mu_2 \leftarrow$ alternative hypothesis, a difference exists.

PROBLEMS

13. Intelligence test on two groups of boys and girls gave the following results :

	Mean	S.D.	N
Girls	75	15	150
Boys	70	20	250

Is there a significant difference in the mean scores obtained by boys and girls ?

Sol :

Let us take the hypothesis that there is no significant difference in the mean scores obtained by boys and girls.

$$\text{S.E.}\left(\bar{X}_1-\bar{X}_2\right) = \sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$$

$$\sigma_1 = 15,\ \sigma_2 = 20,\ n_1 = 150,\ n_2 = 250$$

Substituting the values

$$S.E\left(\bar{X}_1 - \bar{X}_2\right) = \sqrt{\frac{(15)^2}{150} + \frac{(20)^2}{250}} = \sqrt{1.5 + 1.6} = 1.761$$

$$\frac{\text{Difference}}{\text{S.E}} = \frac{75 - 70}{1.761} = 2.84$$

Since the difference is more than 2-58 S.E. (1% level of significance), the hypothesis is rejected. There seems to be a significant difference in the mean scores obtained by boys and girls.

14. A man buys 50 electric bulbs of 'Philips' and 50 electric bulbs of 'HMT'. He finds that 'Philips' bulbs give an average life of 1,500 hours with a standard deviation of 60 hours and 'HMT' bulbs gave an average life of 1,512 hours with a standard deviation of 80 hours. Is there a significant difference in the mean life of the two makes of bulbs?

Sol :

Let us set up the hypothesis that there is no significant difference in the mean life of the two makes of bulbs. Calculating, standard error of difference of means

$$S.E.\ \left(\bar{X}_1 - \bar{X}_2\right) = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$$\sigma_1 = 60,\ n_1 = 50,\ \sigma_2 = 80,\ n_2 = 50$$

$$S.E.\left(\bar{X}_1 - \bar{X}_2\right) = \sqrt{\frac{(60)^2}{50} + \frac{(60)^2}{50}} = \sqrt{\frac{3600 + 6400}{50}} = \sqrt{200} = 14.14$$

Observed difference between the two means = 1512 – 1500 = 12

$$\frac{\text{Difference}}{\text{S.E}} = \frac{1}{14.14} = 0.849$$

Since the difference is less than 2-58 S.E. (1% level of significance), it could have arisen due to fluctuations of sampling. Hence the difference in the mean life of the two makes is not significant.

15. A simple sample of the height of 6.4Q0 Englishmen has a mean of 67-85 inches and a standard deviation of 2-56 inches while a simple sample of heights of 1,600 Austrians has a mean of 68-55 inches and standard deviation of 2-52 inches. Do the data indicate that the Austrians are on the average taller than the Englishmen ? Give reasons for your answer.

Sol :

Let us take the hypothesis that there is no significant difference in the mean height of Englishmen and Austrians.

$$S.E.\left(\bar{X}_1 - \bar{X}_2\right)$$

Since the difference is more than S.E. (at 1% level of significance), the hypothesis is rejected. Hence, the data indicates that the Austrians are on the average taller than the Englishmen.

16. In a survey of buying habits, 400 women shoppers are chosen at random in super market A located in a certain section of Mumbai city. Their average monthly food expenditure is Rs. 250 with a standard deviation of Rs. 40. For 400 women shoppers chosen at random in super market B in another section of the city, the average monthly food expenditure is Rs. 220 with a standard deviation of Rs. 55. Test at 1% level of significance whether the average food expenditure of the two populations of shoppers from which the samples were obtained are equal.

Sol :

Let us take the hypothesis that there is no difference in the average food expenditure of the two populations of shoppers.

S.E. of the difference of means is given by

$$\text{S.E.}\left(\bar{X}_1 - \bar{X}_2\right) = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$n_1 = 400$, $\bar{X}_1 = 250$, $\sigma i = 40$; $n_1 = 400$, $\bar{X}_2 = 220$, $\sigma_2 = 55$

Substituting the values

$$\text{S.E.}\left(\bar{X}_1 - \bar{X}_2\right) = \sqrt{\frac{(40)^2}{400} + \frac{(55)^2}{400}}$$

$$= \sqrt{\frac{1600}{400} + \frac{3025}{400}} = 3.4$$

Difference of Means $= \left(\bar{X}_1 - \bar{X}_2\right)$

$= 250 - 220 = 30$

$$\frac{\text{Difference}}{\text{S.E.}} = \frac{30}{3.4} = 8.82$$

Since the difference is more than 2-58 S.E. (1% level) the hypothesis, is rejected, Hence the average food expenditures of the populations of shoppers are not equal.

4.3.1 Dependent samples

Q11. What are dependent samples? How to test hypothesis with dependent samples.

Ans : **(Imp.)**

In dependent samples, each observation in one sample can be paired with an observation in the other sample.

The following samples could be collected on the same group of units (e.g., individuals):

- At two different points in time (e.g. respond to the same question at two points in time)
- For two different items (e.g. respond to two questions)

These samples can be collected on natural pairings:

Collect the same measurement on relatives

Disease status of two groups, exposed and non-exposed, when they are similar in other respects

Responses from such dependent samples are statistically dependent. Matched pairs are two samples that are statistically dependent. Square tables are tables where the row and column classifications are the same. In these cases you can expect a priori to have dependent samples. Thus, the methods that treat them as independent samples are inappropriate.

4.3.2 Testing for difference between Proportions – Large Sample

Q12. Explain briefly about Hypothesis Concerning one Proportion.

Ans :

Hypothesis concerning One Proportion (Large Sample)

When the value of the sample size n is large, approximation procedures are required. When the value p_0 is very close to 0 or 1, the Poisson

distribution, with parameter $\mu = np_0$ and $\sigma^2 = np_0q_0$ is usually used for large n and is very accurate as long as p_0 is not extremely close to 0 or 1.

If we use the normal approximation, the z- value for testing $p = p_0$ is given by,

$$Z = \frac{x - np_0}{\sqrt{np_0q_0}}$$

Where, x is the number of success in a sample of size n and $q_0 = 1 - p_0$. This statistic can also be written as,

$$Z = \frac{p - P}{\sqrt{pq/n}}$$

Where,

$p = \frac{x}{n}$ = Proportion of success in the sample

P = Actual population of success.

The critical regions for α L.O.S.

Test Alternative Hypothesis Critical Region

(a) Two tailed test $p \neq p_0$ $Z < -Z_{\alpha/2}$ (or) $Z > Z_{\alpha/2}$

(b) Right one tailed test $p > p_0$ $Z > Z_\alpha$

(c) Left one tailed test $p < p_0 Z < -Z_\alpha$

PROBLEMS

17. A manufacturer claimed that atleast 95% of the equipment which the supplied to a factory conformed to specifications. An examination of a sample of 200 pieces of equipment revealed that 18 were faulty. Test his claim at 5% level of significance.

Sol :

Given sample size, n = 200

Number of pieces confirming to specification = 200 – 18 = 182

$\therefore$ p = Proportion of pieces confirming to specifications

$$= \frac{182}{200} = 0.91$$

P = Population proportion = $\frac{95}{100}$ = 0.95

(i) Null Hypothesis H_0 = The proportion of pieces of pieces confirming to specifications

i,e, P = 95%

(ii) Alternative Hypothesis H_1 : P < 0.95 (left - tail test)

(iii) The test statistic is $z = \frac{p-P}{\sqrt{\frac{PQ}{n}}} = \frac{0.91-0.95}{\sqrt{\frac{0.95\times 0.05}{200}}} = \frac{-0.04}{0.0154} = -2.59.$

Since alternative hypothesis is left tailed, the tabulated value of Z at 5% level of significance is 1.645.

Since calculated value of |z| = 2.6 is greater than 1.645, we reject the Null Hypothesis H_0 at 5% level of significance and conclude that the manufacture's claim is rejected.

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

Sol :

Given

n = 1000

p = Sample proportion of rice eaters = $\frac{540}{1000}$ = 0.54

P = Population proportion of rice eaters = $\frac{1}{2}$ = 0.5

$\therefore$ Q = 0.5

(i) Null Hypothesis H_0 : Both rice and wheat are equally popular in the state.

(ii) Alternative Hypothesis H_1 : P $\neq$ 0.5 (two tailed alternative)

(iii) Test statistic is $z = \frac{p-P}{\sqrt{\frac{PQ}{n}}} = \frac{0.54-0.5}{\sqrt{\frac{0.5\times 0.5}{1000}}} = 2.532$

The calculated value of z = 2. 532

The tabulated value of z at 1% level of significance for two - tailed test is 2.58.

Since calculated z < tabulated z, we accept the Null Hypothesis H_0 at 1% level of significance and conclude that both rice and wheat are equally popular in the state.

19. In a big city 325 men out of 600 men were found to be smokers. Does this information support the conclusion that the majority of men in this city are smokers?

Sol :

Given n = 600

Number of smokers = 325

p = sample proportion of smokers = $\frac{325}{600}$ = 0.5417

P = Population proportion of smokers in the city = $\frac{1}{2}$ = 0.5

Q = 1 – P = 1 – 0.5 = 0.5

(i) Null Hypothesis H_0 : The number of smokers and non-smokers are equal in the city

(ii) Alternative Hypothesis : P > 0.5 (Right tailed)

(iii) The Test statistic is $z = \frac{p - P}{\sqrt{\frac{PQ}{n}}} = \frac{0.5417 - 0.5}{\sqrt{\frac{0.5 \times 0.5}{600}}} = 2.04$

$\therefore$ Calculated value of Z = 2.04

Tabulated value of z at 5% level of significance for right tail test is 1.645.

Since, calculated value of z > tabulated value of z, we reject the Null Hypothesis and conclude that the majority of men in the city are smokers.

Q13. Explain briefly about Hypothesis Concerning two Proportion.

Ans : **(Imp.)**

There are many situations in which we wish to test the hypothesis that two proportions are equal. For example, we might want to show evidence that the proportion of engineers in state is equal to the proportion of engineers in another state. A person may decide to give up smoking only if he or she is convinced that the proportion of smokers with lung cancer is more than the proportion of non smokers with lung cancer.

In general, we want to test the null hypothesis that two proportions, or binomial parameters, are equal. That is we wish to test the null hypothesis $H_0 : p_1 = p_2$, against one of the alternatives $p_1 < p_2$, $p_1 > p_2$, or $p_1 \neq p_2$.

Of course, this is equivalent to testing the null hypothesis that $p_1 - p_2 = 0$, against one of the alternatives $p_1 - p_2 < 0$, $p_1 - p_2 > 0$, or $p_1 - p_2 \neq 0$. $p_1 - p_2$ is a random variable on which we base our discussion. Let us suppose that there are two distinct populations A and B. Independent samples of size n_1 and n_2 are selected at random from this two binomial populations and the proportion of success p_1 and p_2 for the two samples is computed.

We know that from previous discussion of construction of confidence intervals for p_1 and p_2, for n_1 and n_2 sufficiently large, the point estimator.

$p_1 - p_2$ was approximately normally distributed with mean

$\mu_{p_1 - p_2} = 0$ and variance $\sigma^2_{p_1 - p_2} = \frac{\hat{p}\hat{q}}{n_1} + \frac{\hat{p}\hat{q}}{n_2}$

Here an unbiased pooled estimate of the population proportion $\hat{p}$ is,

$$\hat{p} = \frac{n_1p_1 + n_2p_2}{n_1 + n_2} = \frac{x_1 + x_2}{n_1 + n_2}$$

Where x_1 and x_2 are the number of success in each of the two samples.

The Z - value for testing $p_1 = p_2$ is determined from the formula,

$$Z = \frac{p_1 - p_2}{\sigma_{p_1+p_2}} = \frac{\frac{x_1}{n_1} - \frac{x_2}{n_2}}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

The critical region for testing the null hypothesis, $p_1 = p_2$ against the number of alternatives at α level of significance.

Null Hypothesis Alternative Hypothesis Critical Region

1. $H_0 : P_1 = P_2 V_s$

   $H_1 : P_1 \neq P_2$

   C.R. : $Z < -Z_{\alpha/2}$

   $Z > Z_{\alpha/2}$

2. $H_0 : P_1 = P_2$

   $H_1 : P_1 > P_2$

   C.R.: $Z > Z_\alpha$

3. $H_0 : P_1 = P_2$

   $H_1 : P_1 < P_2$

   C.R.: $Z < Z_\alpha$

PROBLEMS

20. Random samples of 400 men and 600 women were asked whether they would like to have a flyover near their residence. 200 men and 325 women were in favour of the proposal. Test the hypothesis that proportions of men and women in favour of the proposal are same, at 5% level.

Sol :

Given sample sizes, $n_1 = 400$, $n_2 = 600$

Proportion of men, $p_1 = \frac{325}{600} = 0.5$

Proportion of women, $p_2 = \frac{325}{600} = 0.5416$

(i) Null hypothesis H_0

Assume that there is no significant difference between the option of men and women as far as proposal of flyover is concerned.

(ii) Alternative hypothesis H_1

$p_1 \neq p_2$ (two tailed)

(iii) The test statistic is $z = \frac{p_1 - p_2}{\sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$

[Method of pooling]

Where $p = \frac{n_1p_1 + n_2p_2}{n_1 + n_2}$

$$= \frac{400 \times \frac{200}{400} + 600 \times \frac{325}{600}}{400 + 600}$$

$$= \frac{-0.0416}{0.032} = 0.525$$

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

$\therefore \quad z =$

$$\frac{0.5 - 0.5416}{\sqrt{0.525 \times 0.475\left(\frac{1}{499} + \frac{1}{600}\right)}}$$

$$= \frac{-0.0416}{0.032} = -1.3$$

Thus $|z| = 1.3$

Since $|z| < 1.96$, we accept the null hypothesis H_0 at 5% level of significance.

i.e., there is no difference of opinion between men and women as far as proposal of flyover is concerned.

21. Random samples of 400 men and 200 women in a locality were asked whether they would like to have a bus stop near their residence. 200 men and 40 women in favour of the proposal. Test the significance between the difference of two proportions at 5% level.

Sol :

Let P_1 and P_2 be the population proportions in a locality who favour the bus stop.

Let the Null Hypothesis be $H_0 : P_1 = P_2$

Then the Alternative Hypothesis is $H_1 : P_1 \neq P_2$

Here n_1 = Number of men in I sample

= 400

n_2 = Number of women in II sample = 400

x_1 = Number of men in favour of the proposal = 200

x_2 = Number of women in favour of the proposal = 40

$$\therefore P_1 = \frac{x_1}{n_1} = \frac{200}{400} = \frac{1}{2} \text{ and } p_2 = \frac{x_2}{n_2} = \frac{40}{200} = \frac{1}{5}$$

We have $p = \frac{n_1p_1 + n_2p_2}{n_1 + n_2} = \frac{x_1 + x_2}{n_1 + n_2}$

$$= \frac{200 + 40}{400 + 200} = \frac{240}{600} = \frac{2}{5}$$

$$\therefore q = 1 - p = 1 - \frac{2}{5} = \frac{3}{5}$$

Assuming that H_0 is true, the test statistic is

$$z = \frac{p_1 - p_2}{\sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$= \frac{\frac{1}{2} - \frac{1}{5}}{\sqrt{\frac{2}{5} \times \frac{3}{5}\left(\frac{1}{400} + \frac{1}{200}\right)}}$$

$$= \frac{0.3}{0.0424} = 7.07 > 1.96$$

Since $z > 1.96$, we reject the Null Hypothesis at 5% level of significance and conclude that the difference between the two proportions is highly significant i.e., there is difference between the men and women in their attitude towards the bus stop near their residence.

22. A manufacturer of electronic equipment subjects samples of two completing brands of transistors to an accelerated performance test. If 45 of 180 transistors of the first kind and 34 of 120 transistors of the second kind fail the test, what can he conclude at the level of significance a = 0.05 about the difference between the corresponding sample proportions ?

Sol :

We have $n_1 = 180$, $x_1 = 45$,

$x_2 = 34$, $n_2 = 120$

and $p_1 = \frac{x_1}{n_1} = \frac{45}{180} = 0.25$,

$$p_2 = \frac{x_2}{n_2} = \frac{34}{120} = 0.283$$

$$\therefore p = \frac{n_1p_1 + n_2p_2}{n_1 + n_2} = \frac{x_1 + x_2}{n_1 + n_2}$$

$$= \frac{45 + 34}{180 + 120} = \frac{79}{300} = 0.263$$

$q = 1 - p = 1 - 0.263 = 0.737$

(i) Null Hypothesis H_0 = $p_1 = p_2$ i.e., there is no difference

(ii) Alternative Hypothesis H_1 : $p_1 \neq p_2$ i.e., there is a difference

(iii) Level of significance : $\alpha = 0.05$

(iv) The test statistic is

$$z = \frac{p_1 - p_2}{\sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

[Method of pooling]

$$= \frac{0.25 - 0.283}{\sqrt{(0.263)(0.737)\left(\frac{1}{180} + \frac{1}{120}\right)}}$$

$$= \frac{-0.033}{\sqrt{(0.194)(0.01388)}}$$

$$= \frac{-0.033}{0.0519}$$

$$= -0.6358$$

$$\therefore |z| = 0.6358$$

Since $|z| < 1.96$, we accept the null hypothesis H_0 at 5% level of significance i.e., the difference between the proportions is not significant.

Q14. Explain briefly about Hypothesis Concerning Differences between Proportion.

Ans :

Test of significance for difference of proportions.

Suppose the difference between two population proportion equal to some constant δ, then the test of hypothesis consists of,

Null hypothesis, $H_0 : P_1 - P_2 = \delta$

Alternative hypothesis,

$H_1 : P_1 - P_2 \neq \delta, P_1 - P_2 > \delta$

or $P_1 - P_2 < \delta$

Test Statistic,

$$Z = \frac{(\hat{p}_1 - \hat{p}_2)}{\hat{P}\hat{Q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

$$= \frac{\left(\frac{x_1}{n_1} - \frac{x_2}{n_2}\right) - \delta}{\sqrt{\frac{\frac{x_1}{n_1}\left(1 - \frac{x_1}{n_1}\right)}{n_1} + \frac{\frac{x_2}{n_2}\left(1 - \frac{x_2}{n_2}\right)}{n_2}}}$$

$$P = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{x_1 + x_2}{n_1 + n_2},$$

$$Q = 1 - P$$

Large sample confidence interval for the difference of two proportions

$$= \hat{P}_1 - \hat{P}_2 \pm Z_{\alpha/2}\sqrt{\frac{\hat{P}_1\hat{Q}_1}{n_1} + \frac{\hat{P}_2\hat{Q}_2}{n_2}}$$

Standard error of $\left(\hat{P}_1 - \hat{P}_2\right) = \sqrt{\frac{\hat{P}_1\hat{Q}_1}{n_1} + \frac{\hat{P}_2\hat{Q}_2}{n_2}}$

Maximum error $E = Z_{\alpha/2}\sqrt{\frac{\hat{P}_1\hat{Q}_1}{n_1} + \frac{\hat{P}_2Q_2}{n_2}}$

PROBLEMS

23. A medical researcher testing the effectiveness of a new drug found that 70% of a random sample of 280 patients improved under this drug. In a control group, 140 patients were given a PLACEBO, 50% of these patients improved. Test at a = 0.05, the effectiveness of the new drug.

Sol :

Let the null hypothesis be that there is no change in the patients because of the drug.

$$H_o : \hat{P}_1 = \hat{P}_2 ,$$

$H_1 : \hat{P}_1 \neq \hat{P}_2$

$$Z = \frac{\hat{P}_1 - \hat{P}_2}{\sqrt{\frac{\hat{P}_1\hat{q}_1}{n_1} + \frac{\hat{P}_2\hat{q}_2}{n_2}}}$$

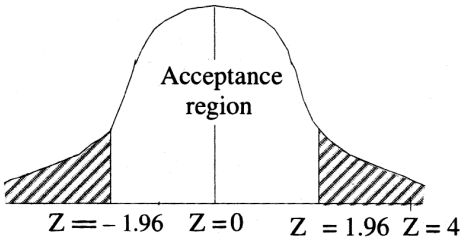
Given data,

$\hat{P}_1 = 0.7; \quad \hat{q}_1 = 1 - \hat{P}_1 = 0.3; \; n_1 = 280$

$\hat{P}_2 = 0.5; \quad \hat{q}_2 = 1 - \hat{P}_2 = 0.5; \; n_2 = 140$

$$\therefore Z = \frac{0.7 - 0.5}{\sqrt{\frac{0.7 \times 0.3}{280} + \frac{0.5 \times 0.5}{140}}} = \frac{0.2}{\sqrt{0.00075 + 0.0018}} = \frac{0.2}{0.05} = 4$$

Rejection region for α = 0.05 is Z > 1.96 and Z < – 1.96, for a two tailed test.



Calculated value of Z = 4

Since, calculated value of Z falls in the rejection region, H_Q is rejected. Hence we may conclude that the proportions of patients improved by new drug and PLACEBO are different. Thus, we can say that the new drug is effective.

UNIT V

Chi-square and Analysis of Variance - chi-square as test of independence, chi-square as a test of goodness of fit, analysis of variance, inferences about a population variance, inferences about two population variances.

Regression and Correlation - Simple Regression - Estimation using regression line, correlation analysis, making inferences about population parameters, limitations, errors and caveats in regression and correlation analysis. Multiple Regression and correlation analysis. Finding multiple regression equations and making inferences about population parameters.

5.1 CHI-SQUARE

Q1. Define Chi-square test.

(OR)

Explain about Chi-square test.

Ans : **(Imp.)**

The χ^2 test (pronounced as chi-square test) is one of the simplest and most widely used non-parametric tests in statistical work. The symbol χ^2 is the Greek letter Chi. The χ^2 test was first used by Karl Pearson in the year 1990. The quantity χ^2 describes the magnitude of the discrepancy between theory and observation. It is defined as :

$$\chi^2 = \Sigma \frac{(O-E)^2}{E}$$

where O refers to the observed frequencies and

E refers to the expected frequencies.

Steps.

To determine the value of χ^2, the steps required are:

(i) Calculate the expected frequencies. In general the expected frequency for any call can be calculated from the following equation :

$$E = \frac{RT \times CT}{N}$$

E = Expected frequency

RT = The row total for the row containing the cell

CT = The column total for the column containing the cell

N = The total number of observations

(ii) Take the difference between observed and expected frequencies and obtain the squares of these differences, i.e., obtain the values of $(O - E)^2$.

(iii) Divide the values of $(O - E)^2$ obtained in step (ii) by the respective expected frequency and obtain the total $\Sigma[(O - E)^2 / E]$. This gives the value of χ^2 which can range from zero to infinity. If χ^2 is zero it means that the observed and expected frequencies completely coincide. The greater the discrepancy between the observed and expected frequencies, the greater shall be the value of χ^2.

The calculated value of χ^2 is compared with the table value of χ^2 for given degrees of freedom at a certain specified level of significance. If at the stated level (generally 5% level is selected), the calculated value of χ^2 is more than the table value of χ^2, the difference between theory and observation is considered to be significant, i.e., it could not have arisen due to fluctuations of simple sampling. If, on the other hand, the calculated value of χ^2 is less than the table value, the difference between theory and observation is not considered as significant, i.e., it is regarded as due to fluctuations of simple sampling and hence ignored.

The computed value of χ^2 is a random variable which takes on different values from sample to sample. That is χ^2 has a sampling distribution just as do the other test statistics.

It should be noted that the value of χ^2 is always positive and its upper limit is infinity. Also since χ^2 is derived from observations, it is a statistic and not a parameter (there is no parameter corresponding to it). The χ^2 test is, therefore, termed nonparametric. It is one of the great advantages of this test that it involves no assumption about the form of the original distribution from which the observations come.

Q2. Explain the assumptions / conditions of χ^2 Test

Ans :

There are five conditions to fulfill for a chi-square test,

1. Sample observation data must be independent of each other.
2. Random sampling from specified population to give sample data.
3. Data should not be in percentage or ratio form but original units to make comparison easy.
4. Sample size should have atleast 50 observations.

Q3. Explain the uses of Chi-square distribution.

Ans :

Uses

1. A. chi-square statistic can be used to test research questions involving cross-tabulated categorical variables.
2. An overall chi-square statistic is computed by summing the individual cell values (chi-squares) in a cross-tabulated table.
3. The degrees of freedom for a cross-tabulated table are row minus one times column minus one, i.e., df = (r – 1) (c – 1).
4. The chi-square test of independence can be used for any number of rows and columns, as long as the expected cell frequency is greater than five.
5. A chi-square test of independence is used to determine whether or not the rows and columns are independent (null hypothesis).
6. If the null hypothesis is true, it is still possible that the chi-square test could lead to a rejection of the null hypothesis (Type I error).
7. If the null hypothesis is false, it is still possible that the chi-square test could lead to retaining the null hypothesis (Type II error).
8. The ratio of each cell value to the overall chi-square value provides a variance accounted for interpretation of how much each cell contributed to the overall chi-square value.

5.1.1 Chi-square as Test of Independence

Q4. "Chi-square act has a Test of Independence". Elaborate.

Ans :

The Chi-Square Test of Independence determines whether there is an association between categorical variables (i.e., whether the variables are independent or related). It is a nonparametric test.

This test is also known as Chi-Square Test of Association.

This test utilizes a contingency table to analyze the data. A contingency table (also known as a cross-tabulation, crosstab, or two-way table) is an arrangement in which data is classified according to two categorical variables. The categories for one variable appear in the rows, and the categories for the other variable appear in columns. Each variable must have two or more categories. Each cell reflects the total count of cases for a specific pair of categories.

5.1.2 Chi-square as a Test of Goodness of Fit

Q5. Explain briefly about test for goodness of fit?

Ans : **(Imp.)**

One of the very popular applications of χ^2 test is test of goodness of fit. It enables us to ascertain how the theoretical distributions such as Binomial, Poisson, Normal etc. can fit into empirical

distributions obtained from sample data. When an ideal frequency curve whether normal or some other type is fitted to the data, we are interested out how well this curve fits with the observed facts.

A test of the concordance (goodness of fit) of the two can be made just by inspection, but such a test is obviously inadequate. Precision can be secured by applying the χ^2 test. The following are the steps in testing the goodness of fit :

1. Null and alternative hypotheses are established, and a significance level is selected for rejection of the null hypothesis.
2. A random sample of independent observations is drawn from a relevant statistical population.
3. A set of expected or theoretical frequencies is derived under the assumptions that the null hypothesis is true. This generally takes the form of assuming that a particular probability distribution is applicable to the statistical population under consideration.
4. The observed frequencies are compared with the expected, or theoretical frequencies.
5. If the calculated value of χ^2 is less than the table value at a certain level of significance (generally 5% level) and for certain degrees of freedom the fit is considered to be good, i.e., the divergence between the actual and expected frequencies is attributed to random fluctuations of sampling.

On the other hand, if the calculated value of χ^2 is greater than the table value, the fit is considered to be poor, i.e., it cannot be attributed to fluctuations of sampling rather it is due to the inadequacy of the theory to fit the observed facts.

Goodness of Fit

χ^2 test help us to find out how well the assumed theoretical distribution fit to the observed data. When some theoretical distribution is fitted to the given data, the statistician or managers will be interested in knowing as to how this distribution fits with the observed data.

This method of χ^2 test helps in answering this question.

If the calculated value of χ^2 is less than the table value at a certain level of significance, the fit is considered to be good one i.e., divergence between the observed and expected frequencies is attributable to fluctuations of sampling. But if the reverse occurs, the fit is not considered to be a good one. In short,

$\chi^2_{cal} < \chi^2_{table} \Rightarrow$ Good fit

$\chi^2_{cal} < \chi^2_{table} \Rightarrow$ Not a good fit.

If = 1, 2, ... n is a set of observed (experimental) frequencies and E_i (i = 1, ... n) is the corresponding set of theoretical frequencies then $\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$ with the condition that, $\sum_{i=1}^{n} O_i = \sum_{i=1}^{n} E_i = N$ = Total frequency follows, χ^2 - Distribution with (n – 1) d.o.f.

Steps for Test of Goodness of Fit

1. Null hypothesis : Good fit exists between the theoretical distribution and given data.
2. Alternative hypothesis : No good fit.
3. Level of significance is α.
4. Critical region : Reject null hypothesis if $\chi^2 > \chi^2_a$ with v d.o.f. i.e., theoretical distribution is a poor fit.
5. Computations : $\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$
6. **Decision :** Accept null hypothesis, if $\chi^2 < \chi^2_\alpha$, i.e., the theoretical distribution is a good fit to the data.

PROBLEMS

1. Assume that Air Ticket reservations from Delhi to Gulf are uniformly distributed during all days in the winter season. To determine whether it is uniform we have selected a random sample of reservation lists for 10 days.

The following information is drawn from the list.

Distribution of actual number of reservations.

Sl. No.	No. of Reservations
1	65
2	80
3	100
4	98
5	75
6	80
7	82
8	70
9	60
10	90

Test the validity of assumption using chi-square test.

Sol :

(i) Null Hypothesis

Air ticket reservations from Delhi to Gulf are uniformly distributed during all days in winter season.

Alternate Hypothesis

Air ticket reservations are not uniformly distributed during all days of winter season. A sample of 800 reservations for 10 days is given.

Therefore, expected reservation for each day is $\frac{800}{10} = 80$

(ii) Computing Test Statistic χ^2

Sl.No.	O	E	O – E	$(O-E)^2$	$(O-E)^2/E$
1	65	80	– 15	225	2.8125
2	80	80	0	0	0
3	100	80	20	400	5
4	98	80	18	324	4.05
5	75	80	-5	25	0.3125
6	80	80	0	0	0
7	82	80	2	4	0.05
8	70	80	– 10	100	1.25
9	60	80	– 20	400	5
10	90	80	10	100	1.25
	800			$\Sigma\left(\frac{(O-E)^2}{E}\right)$	= 19.725

$$\chi^2 = \Sigma\left[\frac{(O-E)^2}{E}\right] = 19.725$$

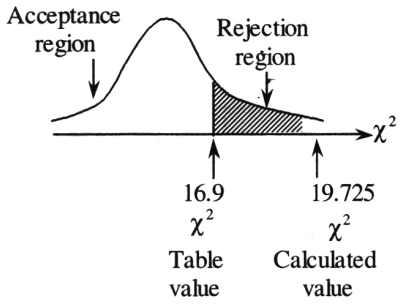
(iii) Level of significance, $\alpha = 0.0.5$

Degrees of freedom, d.f = n – 1 = 10 – 1 = 9

(iv) Table value χ^2 for 9 d.f at 5% level of significance is 16.9

(v) Decision

The calculated value of χ^2 (19.725) is greater than table value 16.9. The null hypothesis is rejected. Hence, it can be said that air reservations from Delhi to Gulf are not uniformly distributed during all days in winter season.



Figure

2. A survey of 320 families with 5 children each revealed the following distribution :

Number of Boys	5	4	3	2	1	0
Number of Girls	0	1	2	3	4	5
Number of Families	14	56	110	88	40	12

Is this result consistent with the hypothesis is that male and female births are equally probable?

Sol :

Null Hypothesis (H_0) : Male and Female births are equally probable against

Alternative Hypothesis (H_1) : Male and Female births are not equally probable.

This assumptions of H_0 takes us to the probability of a male birth is p = $^1/_2$ and the underlying distribution is Binomial distribution. Since the birth can be a Male or Female the dichotomous classification.

Fix a = 5%.

Test Statistic : $\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$ ~ a Chi Square Distribution with (k–1) D.F., where O_i refers to the observed frequencies and E_i refers to the expected frequencies. The probability of x male births in a family of 5 is given by

$P(x) = {}^5C_x\ P^x\ q^{5-x}$; for $x = 0, 1, 2, 3, 4, 5,$

$= {}^5C_x\ (1/2)^5$ [since $p = q = 1/2$]

To get the expected frequencies, multiply P(x) by the total number N = 320.

X	P(x)	Expected Frequency = N × P(x)
0	${}^5C_0\ (1/2)^5 = 1/32$	320 × 1/32 = 10
1	${}^5C_1\ (1/2)^5 = 5/32$	320 × 5/32 = 50
2	${}^5C_2\ (1/2)^5 = 10/32$	320 × 10/32 = 100
3	${}^5C_3\ (1/2)^5 = 10/32$	320 × 5/32 = 100
4	${}^5C_4\ (1/2)^5 = 5/32$	320 × 5/32 = 50
5	${}^5C_5\ (1/2)^5 = 1/32$	320 × 1/32 = 10

Arranging the observed and expected frequencies in the following table and calculating c^2.

O_i	E_i	$(O_i - E_i)^2$	$(O_i - E_i)^2/E_i$
14	10	16	1.60
56	50	36	0.72
110	100	100	1.00
88	100	144	1.44
40	50	100	2.00
12	10	4	0.40
			$\Sigma(O_i - E_i)^2 / E_i = 7.16$

Table Value

At 5% LOS (selected), the table value of c^2 for v = 6–1 = 5 at 5% LOS in two tailed test is between (0.831209), 12.832492) (found from the tables in Appendix).

Conclusion

Since the computed value of $c^2 = 7.16$ is falling within the acceptance region (limits obtained from the tables) for 5 D.F. at 5% LOS, we may accept H_0 and conclude that the male and female births are equally probable.

3. The following results are obtained when a dice is thrown 132 times :

Number Turned up	1	2	3	4	5	6
Frequency	16	20	25	14	29	28

Test the hypothesis is that the dice is unbiased.

Sol :

1) **Hypothesis (H_0):** The dice is unbiased

Against alternative Hypothesis (H_1): These dice is biased. This assumption of H_0 takes us to the probability of each face as $\left(\frac{1}{6}\right)$.

2) Fix ∞ = 2%

3) Test Statistic : $\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$ ~ a Chi-Square Distribution with (k–1) D.F., where O_i refers to the observed frequencies and E_i refers to the expected frequencies. The expected frequencies for each face is $132 \times \frac{1}{6} = 22$, with the assumption that the dice is unbiased.

Computation of the value of χ^2

Number turning up O_i	Frequency Observed E_i	Expected	$O_i - E_i$	$(O_i - E_i)^2$	$(O_i - E_i)^2/E_i$
1	16	22	–6	36	36/22
2	20	22	–2	4	4/22
3	25	22	3	9	9/22
4	14	22	–8	64	64/22
5	29	22	7	49	49/22
6	28	22	6	36	36/22
				Total χ^2	198/22 = 9

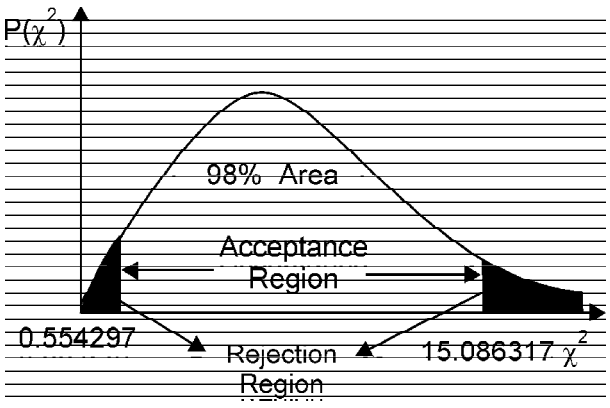


Table Value

At 2% LOS (selected), the table value of c^2 for n = 6 – 1 = 5 D.F. in two tailed test is between (0.554297, 15.086317) (found from the tables is Appendix).

Conclusion

Since the computed value of $c^2 = 9$ is falling within the acceptance Region (limits obtained from the tables) for 5 D.F. at 2% LOS, we may accept H_0 and conclude that the Dice is an unbiased one.

5.2 Analysis of Variance

Q6. What is ANOVA? What are its assumptions and applications?

Ans : **(Imp.)**

ANOVA

The variance test is also known as ANOVA. ANOVA is the acronym for Analysis of Variance. Analysis of variance is a statistical technique specially designed to test whether the means of more than two quantitative populations are equal i.e., to make inferences about whether those samples are drawn from the populations having the same mean.

The test is called 'F' test as it was developed by R.A Fisher in 1920's. The test is conducted in situations where we have three or more to consider, at a time an alternative procedure (to t-test) needed for testing the hypothesis that all samples could likely be drawn from the same population.

Example

Five fertilizers are applied to four plots, each of wheat and yield of wheat on these plots is given. We are interested in finding out whether the effects of these fertilizers on the yields are significantly different or, in other words, whether the samples have come from the same population. ANOVA answers this question.

Assumptions

Analysis of variance test is based on the test statistic F (or variance ratio).

It is based on the following assumptions,

(i) Observations are independent.

(ii) Each sample is drawn randomly from a normal population as the sample statistics reflect the characteristic of the population.

(iii) Variance and means are identical for those population from which samples have been drawn.

Applications

The applications of ANOVA are as follows,

1. Anova is used in education, industry, business, psychology fields mainly in their experiment design.
2. Anova helps to save time and money as several population means can be compared simultaneously.
3. Anova is used to test the linearity of the fitted regression line and correlation ratio, significance test statistic of anova

$= F(r - 1, n - r)$.

5.2.1 Inferences about a Population Variance

Q7. Explain briefly about one ways Anova.

Ans :

In these classification the data is classified according to only one criteria i.e., It includes only one factor.

Steps involved in ANOVA One Way

(i) Null Hypothesis $H_0 = \mu_1 = \mu_2 = \mu_3 = ... = \mu_k$ (where means are equal)

(ii) Alternative Hypothesis $H_1 = \mu_1 \neq \mu_2 ... \neq \mu_k$ (when means are not equal)

Arithmetic mean and drawn from the means of population from which "K" samples are drawn which are equal to another.

Step 1

(i) Calculation variance between samples

(ii) Calculation of grand average $\bar{\bar{X}}$

$$\bar{\bar{X}} = \frac{\bar{X}_1 + \bar{X}_2 + \bar{X}_3}{N}$$

(iii) Take the difference between means of variance samples of grand average.

(iv) Square the deviations and obtain total which will give some of squares between samples.

(v) Divide the total by degrees of freedom K = No. of samples $\boxed{V = K - 1}$

Step 2

(i) Calculate variance within samples.

(ii) Calculation of mean value of samples $\bar{X}_1, \bar{X}_2, \ldots$

(iii) Take the deviation of variance items in α samples from mean values.

(iv) Square the derivations and obtain total which give sum of squares within samples.

(v) Divide the total by degrees of freedom

$$\boxed{V = N - K}$$

N = No. of observations

K = Refers to the no. of samples

Step 3

Calculation of "F" Ratio

$$\text{"F"} = \frac{S_1^2}{S_2^2}$$

Step 4

Compare the calculated value of "F" with 5% level of significance.

- If $F_{cal} > F_{tab} \rightarrow$ Difference between sample means are significant.
- If $F_{cal} < F_{tab} \rightarrow$ Difference between sample means are not significant.

Step 5

Sources of Variations	SS (Sum of Squares)	V = Degress of Freedom	MS Mean Squares	"F" Ration
Between Samples	SSC	$V_1 = C - 1$	$MSC = \frac{SSC}{C-1}$	
Within Samples	SSE	$V_2 = M - C$	$MSE = \frac{SSE}{N-C}$	$\frac{MSC}{MSE}$
Total	SST	n – 1		

SSC = Sum of Squares between samples (columns)

SSE = Sum of Squares within samples (rows)

SST = Total Sum & Squares of Variations

MSC = Mean sum of squares between samples

MSE = Mean sum of squares within samples.

PROBLEMS

4. Test whether the significance of possible variation in performance in a certain test between the grammar schools of a City A common test is given to number of students taken of random from the senior Vth Class of four schools.

		Schools		
S.No	A	B	C	D
1	8	12	18	13
2	10	11	12	9
3	12	9	16	12
4	8	14	6	16
5	7	4	8	15

Sol :

Given

No. of samples (k) = 4 samples (A, B, C, D)

Hypothesis

Null Hypothesis : There is no significant difference between schools

$H_0 = \mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4$

Alternative Hypothesis : There is significant difference between schools

$H_1 = \mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4$

Step 1: Calculation of Variance between Samples

Calculation of $\bar{X}_1, \bar{X}_2, \bar{X}_3, \bar{X}_4$

X_1	X_2	X_3	X_4
8	12	18	13
10	11	12	9
12	9	16	12
8	14	6	16
7	4	8	15
$\bar{X}_1 = 9$	$\bar{X}_2 = 10$	$\bar{X}_3 = 12$	$\bar{X}_4 = 13$

Calculation of grand mean $\bar{\bar{X}}$

$$\bar{\bar{X}} = \frac{\bar{X}_1 + \bar{X}_2 + \bar{X}_3 + \bar{X}_4}{N}$$

$$\bar{\bar{X}} = \frac{9 + 10 + 12 + 13}{4}$$

$\boxed{\bar{\bar{X}} = 11}$

Calculation of variance between samples

Sample (A) $(\bar{X}_1 - \bar{\bar{X}})^2$	Sample (B) $(\bar{X}_2 - \bar{\bar{X}})^2$	Sample (C) $(\bar{X}_3 - \bar{\bar{X}})^2$	Sample (D) $(\bar{X}_4 - \bar{\bar{X}})^2$
$(9-11)^2=4$	$(10-11)^2=1$	$(12-11)^2=1$	$(13-11)^2=4$
$(9-11)^2=4$	$(10-11)^2=1$	$(12-11)^2=1$	$(13-11)^2=4$
$(9-11)^2=4$	$(10-11)^2=1$	$(12-11)^2=1$	$(13-11)^2=4$
$(9-11)^2=4$	$(10-11)^2=1$	$(12-11)^2=1$	$(13-11)^2=4$
$(9-11)^2=4$	$(10-11)^2=1$	$(12-11)^2=1$	$(13-11)^2=4$
20	5	5	5

Sum of the squares of between samples

$= 20 + 5 + 5 + 20$

$= 50$

Mean sum of the squares of between samples

$$= \frac{50}{K-1}$$

$$= \frac{50}{4-1}$$

$$= \frac{50}{3}$$

$= 16.6$

Step 2 : Calculation Variance with in Samples

Sample (A)		Sample (B)		Sample (C)		Sample (D)	
X_1	$(X_1-\bar{X}_1)^2$	X_2	$(X_2-\bar{X}_2)^2$	X_3	$(X_3-\bar{X}_3)^2$	X_4	$(X_4-\bar{X}_4)^2$
8	1	12	4	18	36	13	0
10	1	11	1	12	0	9	16
12	9	9	1	16	16	12	1
8	1	14	16	6	36	16	9
7	4	4	36	8	16	15	4
	16		58		104		30

Sum of the square within samples

= 16 + 58 + 104 + 30

= 208

Mean sum of the squares of within samples

$$= \frac{208}{M - K}$$

M = Total no. of observations

K = No. of samples

$$= \frac{208}{20 - 4} = \frac{208}{16} = 13$$

Step 3 : Calculation of "F" Ratio

Sources of Variations	Sum of Squares	Degrees of Freedom	MS	"F" Ration
Between Samples	50	$V_1 = C - 1$ $V_1 = 4 - 1$ $V_1 = 3$	$MSC = \frac{SSC}{C - 1}$ 16.6	
Within Samples	208	$V_2 = M - C$ $V_2 = 20 - 4$ $V_2 = 16$	$MSE = \frac{SSC}{M - C}$ 13	$F = \frac{MSC}{MSE}$ $F = 1.27$
Total	258	19		

Step 4 : Acceptance & Reject

(V_1 = 3, V_2 = 16 $t_{0.05}$ = 3.24)

$F_{cal} < F_{tab}$

1.24 < 3.24

∴ Accept H_0

∴ There is no significance difference between schools.

5. Four machines A, B, C and D are used to produce a certain kind of cotton fabrics. Samples of size 4 with each unit as 100 square meters are selected from the outputs of the machines at random and the number of flaws in each 100 square meters show the following result.

A	B	C	D
8	6	14	20
9	8	12	22
11	10	18	25
12	4	9	23

Do you think that there is a significant difference in the performance of the four machines?

Sol :

Let the null hypothesis be that there is no significant difference in the performance of the four machines i.e., $\mu_0 = \mu_1 = \mu_{11} = \mu_{111}$.

A	B	C	D	Total
8	6	14	20	48
9	8	12	22	51
11	10	18	25	64
1 2	4	9	23	48
40	28	53	90	GT = 211

Correction Factor (C.F) = $\frac{(GT)^2}{N}$ = N = 16

C.F. = $\frac{(211)^2}{16}$ = 2782.56

Total Sum of Squares (TSS)

$= \Sigma_i \Sigma_j X_{ij}^2 - C.F$

$= [(8)^2 + (6)^2 + (14)^2 + (20)^2 + (9)^2 + (8)^2 + (12)^2 + (22)^2 + (11)^2 + (10)^2 + (18)^2.$
$+ (25)^2 + (12)^2 + (4)^2 + (9)^2 + (23)^2] - 2782.56$

$= [64 + 36 + 196 + 400 + 81 + 64 + 144 + 484 + 121 + 100 + 324 + 625 + 144$
$+ 16 + 81 + 529] - 2782.56$

$= 3409 - 2782.56 = 626.44$

Sum of squares between the samples,

$$(SSB) = \sum_j \frac{T_j^2}{n_j} - \frac{(GT)^2}{N}$$

$$= \left[\frac{(40)^2}{4} + \frac{(28)^2}{4} + \frac{(53)^2}{4} + \frac{(90)^2}{4}\right] - 2782.56$$

$= [400 + 196 + 702.25 + 2025] - 2782.56$

= 3323.25 – 2782.56

= 540.69

Sum of squares within samples,

SSW = TSS – SSB

= 626.44 – 540.69

= 85.75

ANOVA Table

Sources of Variation	Degree of Freedom	Sum of Squares	Mean Square	F-Ratio
Between Samples	(k–1) = (4–1) = 3	540.69	$\frac{540.69}{3} = 180.23$	$\frac{180.23}{7.14} = 25.22$
Within samples	(n–k) = 16–4 = 12	85.75	$\frac{85.75}{12} = 7.14$	
Total	(n–1) = 16–1 = 15			

F-ratio$_{(3, 12)}$ = calculated = 25.22

F-ratio from table $V_1 = 3$ and $V_2 = 12$ at 5% level of significance = 3.49

Since $F_{(3, 12)}$ calculated > $F_{(3, 12)}$ table value we reject H_0 which means that there is a significant difference between the performance of four machines.

6. Three different machines are used for a production. On the basis of the outputs, test whether the machines are equally effective.

OUTPUTS

Machine 1	Machine 2	Machine
10	9	20
5	7	16
11	5	10
10	6	4

Sol :

Let μ_1, μ_2, μ_3 be the population means of production by three machines 1,2 and 3 respectively.

Null Hypothesis $H_0 : \mu_1 = \mu_2 = \mu_3$ i.e,the machines are equally effective.

Alternative Hypothesis $H_1 : \mu_1, \mu_2, \mu_3$ are not equal i.e., the machines are not all equally effective.

The calculations for sample means, the variances between and within the samples are shown below:

CALCULATIONS FOR SAMPLE MEANS

Machine 1 (Sample 1) X_1	Machine 2 (Sample 2) X_2	Machine (Sample 3) X_3
10	9	20
5	7	16
11	5	10
10	6	4
$36 = \Sigma X_1$	$27 = \Sigma X_2$	$60 = \Sigma X_3$

$\bar{X}_1 = \frac{\Sigma X_1}{n_1} = \frac{36}{4} = 9,\ \bar{X}_2 = \frac{\Sigma X_2}{n_2} = \frac{27}{4} = 6.75,\ \bar{X}_3 = \frac{\Sigma X_3}{n_3} = \frac{60}{4} = 15$

Grand Mean, $\bar{X} = \frac{\bar{X}_1 + \bar{X}_2 + \bar{X}_3}{3} = \frac{9 + 6.75 + 15}{3} = 10.25$

Now SSB = Sum of the Squares between the Samples

$= \Sigma n_i \left(\bar{X}_i - \bar{X}\right)^2$

$= 4(9 - 10.25)^2 + 4(6.75 - 10.25)^2 + 4(15 - 10.25)^2$ [$\because n_1 = n_2 = n_3 = 4$]

$= 6.25 + 49 + 90.25 = 145.5$

and v_1 = degrees of freedom $= k - 1 = 3 - 1 = 2$

$\therefore$ MSB = Mean Square between the samples

$= \frac{SSB}{v_1} = \frac{145.5}{2} = 72.75$

CALCULATIONS FOR SSW

Sample 1		Sample 2		Sample 3	
X_1	$\left(X_1 - \bar{X}_1\right)^2$	X_2	$\left(X_2 - \bar{X}_2\right)^2$	X_3	$\left(X_3 - \bar{X}_3\right)^2$
10	1	9	5.0625	20	25
5	16	7	0.0625	16	1
11	4	5	3.0625	10	25
10	1	6	0.5625	14	1
Total	$22 = \Sigma\left(X_1 - \bar{X}_1\right)^2$		$8.75 = \Sigma\left(X_2 - \bar{X}_2\right)^2$		$52 = \Sigma\left(X_3 - \bar{X}_3\right)^2$

SSW = Sum of the Squares within the samples

$$= \Sigma\left(X_1 - \bar{X}_1\right)^2 + \Sigma\left(X_2 - \bar{X}_2\right)^2 + \Sigma\left(X_3 - \bar{X}_3\right)^2$$

$$= 22 + 8.75 + 52 = 82.75$$

and v_2 = degrees of freedom = N – k = 12 – 3 = 9

$\therefore$ MSW = Mean Square within the samples

$$= \frac{SSW}{v_2} = \frac{82.75}{9} = 9.194$$

ANOVA TABLE

Source of Variation (SV)	Sum of Squares (SS)	Degree of Freedom (d.f)	Mean Squares (MS)	Test statistic (F - ratio of variance)
Between Samples (Columns)	145.5	2	72.75	
Within Samples	82.75	9	9.94	$F = \frac{72.75}{9.194} = 7.913$
Total	228.25	11	–	–

From the table given at the book, the value of F for $v_1 = 2$ and $v_2 = 9$ at 5% level is 4.26. We see that the calculated value 7.913 of F is greater than the tabulated value 4.26. Hence, we reject the Null Hypothesis at 5% level and conclude that the three machines are not equally effective.

Q8. Explain briefly about two way ANOVA with and without interaction?

Ans :

Two way classification/two factor ANOVA is defined where two independent factors have an effect on the response variable of interest.

Example : Yield of crop affected by type of seed as well as type of fertilizer.

Procedure

(a) Calculate the variance between columns,

$$SSC = \sum_{j=1}^{c} \frac{T_j^2}{n_j} - \frac{T^2}{N}$$

(b) Calculate the variance between rows,

$$SSR = \sum_{i=1}^{r} \frac{T_i^2}{n_i} - \frac{T^2}{N}$$

(c) Compute the total variance,

$$SST = \sum X_{ij}^2 - \frac{T^2}{N}$$

(d) Calculate the variance of residual or error,

$$SSE = TSS - (SSC + SSR)$$

(e) Divide the variances of between columns, between rows and residue by their respective degrees of freedom to get the mean squares.

(f) Compute F ratio as follows,

F-ratio concerning variation between columns,

$$= \frac{\text{Mean square between columns}}{\text{Mean squares of residual}}$$

F-ratio concerning variation between rows,

$$= \frac{\text{Mean square between rows}}{\text{Mean squares of residual}}$$

(g) Compare F-ratio calculated with F-ratio from table,

If F-ratio (calculated) < F-ratio (table), H_1 accepted,

If F-ratio (calculated) $\geq$ F-ratio (table), H_0 rejected,

H_1 accepted $\Rightarrow$ no significant differences

H_0 rejected $\Rightarrow$ significant differences

Two-Way ANOVA with Interaction

Under two-way ANOVA with interaction, the total sum of squares SST is divided into four components, which are as follows,

1. The Sum of Squares for factor 'A' (SSA)
2. The Sum of Squares for factor 'B' (SSB)
3. The Sum of Squares for the interaction between two factors 'SSAB'.
4. The Error of Sum of Squares (SSE).

These factors can be represented as,

$$SST = SSA + SSB + SSAB + SSE$$

The main purpose of using two-way ANOVA with interaction is to understand the relationship between factors 'A' and 'B'. Such relationship will help to find out the impact, effect or influence of factor 'A' on factor 'B' and factor 'B' on factor 'A'.

Two-Way ANOVA without Interaction

Under two-way ANOVA without interaction, the total variability of data is divided into three components, which are as follows,

1. Treatment i.e., factor 'A'
2. Block i.e., factor 'B'
3. Chance.

However, the term 'block' refers to a matched group of observations from each population. When units of each block are assigned randomly to each treatment then the design of such experiment is referred as randomized block design.

Note :

Two factors are said to interact if the difference between levels (treatments) of one factor depends on the level of the other factor. Factors that do not interact are called additive.

A combination of a treatment from one factor with a treatment from another factor results in an interaction.

An interaction between two factors exists when for atleast one combination of treatments, the effect of combination is not additive.

ANOVA Table for Two-way Classified Data with m-Observation Per Cell

Sources of Variation	Degree of Freedom	S.S	M.S.S	Variance Ratio F
Factor A	p – 1	S_A^2	$S_A^2 = \frac{S_A^2}{p-1}$	$F_A = \frac{S_A^2}{S_E^2}$
Factor B	q – 1	S_B^2	$S_B^2 = \frac{S_B^2}{q-1}$	$F_B = \frac{S_B^2}{S_E^2}$
Interaction AB	(p–1) (q–1)	S_{AB}^2	$F_{AB} = \frac{S_{AB}^2}{S_E^2}$	
Factor AB	pq (m – 1)	S_E^2	$S_E^2 = \frac{S_E^2}{pq(m-1)}$	
Total	pqm – 1			

Remark

The calculation of various sum of squares is facilitate to a great extent by the use of following formulae,

$$C.F = \frac{G^2}{pqm} = \frac{T^2}{pqm}$$

$$TSS = \sum_i \sum_j \sum_k x_{ij}k^2 - C.F = RSS - CF = \sum_i T_i^2$$

$$S_A^2 = \frac{i}{M} CF$$

$$S_B^2 = \frac{j}{M} CF$$

$$S_{AB}^2 = \sum \sum T_{ij}^2$$

SS due to Means (SSM)

$$= \frac{ij}{mp} CF$$

$$S_{AB}^2 = SSM - S_A^2 - S_B^2$$

$$S_E^2 = TSS - S_A^2 - S_B^2 - S_{AB}^2$$

Hypothesis Tests in Two-way ANOVA

- **Factor A Test :** Hypothesis is designed to determine whether there are any factor A main effects. Null Hypothesis true if and only if there are no differences in means due to different treatments (population) of factor A.
- **Factor B Test :** Hypothesis is test designed to detect factor B main effects. Null hypothesis is true if and only if there are no differences in means due to different treatments (populations) of factor B.
- **Test for AB Interactions :** Test for existence of interactions between levels of the two factors Null hypothesis is true if and only if there are no two way interactions between levels of factor A and levels of factor B, means factor effects are additive for two way ANOVA.

PROBLEMS

7. Four different drugs have been developed for a certain disease. These drugs are used under three different environments. It is assumed that the environment might affect efficacy of drugs. The number of cases of recovery from the disease per 100 people who have taken the drugs is tabulated as follows :

Environment	Drug A1	Drug A2	Drug A3	Drug A4
I	19	8	23	8
II	10	9	12	6
III	11	10	13	16

Test whether the drugs differ in their efficacy to treat the disease, also whether there is any effect of environment on the efficacy of disease.

Sol :

Null Hypothesis

H_o = There is no significant difference in the efficacy of drugs to treat the disease.

H_o = There is no significant effect of environment on the efficacy of disease.

Environment	Drug				Total
	A_1	A_2	A_3	A_4	
I	19	8	23	8	58
II	10	9	12	6	37
III	11	10	13	16	50
Total	**40**	**27**	**48**	**30**	**GT=145**

$$\text{Correction Factor, (CF)} = \frac{(\text{Grand Total})^2}{N}$$

$$= \frac{(145)^2}{12} = 1752.08$$

Total Sum of Squares (TSS)

$$(TSS) = \sum_i \sum_j X_{ij}^2 - C.F$$

$$= [(19)^2 + (8)^2 + (23)^2 + (8)^2 + (10)^2 + (9)^2 + (12)^2 + (6)^2 + (11)^2 + (10)^2 + (13)^2 + (16)^2] - C.F$$

$$= 2025 - 1752.08$$

$$\therefore \quad TSS = 272.92$$

Sum of Squares Between Drugs (Column)

$$SSC = \sum_j \frac{T_j^2}{n_j} - \frac{(GT)^2}{N}$$

$$= \left[\frac{(40)^2}{3} + \frac{(27)^2}{3} + \frac{(48)^2}{3} + \frac{(30)^2}{3}\right] - 1752.08$$

$$= (533.33 + 243 + 768 + 300) - 1752.08$$

$$= 1844.33 - 1752.08$$

$$\therefore \quad SSC = 92.25$$

$$\text{Degree of freedom (r)} = (C - 1)$$

$$= (4 - 1)$$

$$= 3$$

Sum of Squares Between Environment (Rows)

$$SSR = \sum_i \frac{T_i^2}{n_i} = \frac{(GT)^2}{N}$$

$$= \left[\frac{(58)^2}{4} + \frac{(37)^2}{4} + \frac{(50)^2}{4}\right] - C.F$$

$$= (841 + 342.25 + 625) - 1752.08$$

$$= 1808.25 - 1752.08$$

$$\therefore \quad SSR = 56.17$$

Degree of freedom,

$$V_m = (r - 1) - (3 - 1) - 2$$

Residual = Total sum of squares – (Sum of squares between columns + Sum of squares between rows)

= TSS – (SSC + SSR) = 272.92 – (92.25 + 56.17)

= 272.92 – 148.42 = 124.5

ANOVA TABLE

Sources of variation	Sum of squares	Degrees of Freedom	Means squares	Variance Ratio (F)
Between Drugs	92.25	(C–1) = (4–1)=3	$\frac{92.25}{3}=30.75$	$F_{(3.6)}=\frac{30.75}{20.75}=1.48$
Between Environment	56.17	(r–1) = (3–1)=2	$\frac{56.17}{2}=28.09$	$F_{(2,6)}=\frac{28.09}{20.75}=1.354$
Residual or Error (e)	124.5	(C–1) (r–1)=3×2=6	$\frac{124.5}{6}=20.75$	
Total	**272.92**	**(12 – 1) = 11**		

[**Note:** As level of significance is not given in the problem, assume 5% level of significance]

Critical value of $F_{0.05}$	Computed value of F
Drugs at V_0 (3,6) = 4.76	1.48
Environment at V_m(2,6) = 5.14	1.354

Table values are calculated as per 5% level of significance.

Decision

1. Drugs

Since the calculated value of F(1.48) is less than the table value (4.76), null hypothesis is accepted. Hence, there is no significant difference in the efficacy drugs.

2. Environment

Since the calculated value of F(1.354) is less than the table Value (5.14), null hypothesis is accepted. Hence, there is no affect of environment on the efficacy of disease.

8. Suppose that we are interested in establishing the yield producing ability of four types of soya beans A, B, C and D. We have three blocks of land X,Y and Z which may be different in fertility. Each block of land is divided into four plots and the different types of soya beans are assigned to the plots in each block by a random procedure. The following results are obtained:

Soya Bean

Block	Type A	Type B	Type C	Type D
X	5	9	11	10
Y	4	7	8	10
Z	3	5	8	9

Test whether A,B,C and D are significantly different.

Sol :

Null Hypothesis

H_o : There is no significant difference between A,B,C and D.

Soya bean

Block	Type A	Type B	Type C	Type D	Total
X	5	9	11	10	35
Y	4	7	8	10	29
Z	3	5	8	9	25
Total	12	21	27	29	GT = 89

$$\text{Correction Factor (CF)} = \frac{(\text{Grand Total})^2}{N} = \frac{(89)^2}{12}$$

$$= 660.08$$

Total Sum of Squares (TSS)

$$= \sum_i \sum_j X_{ij}^2 - \frac{(GT)^2}{N}$$

$$= [(5)^2 + (9)^2 + (11)^2 + (10)^2 + (4)^2 + (7)^2 + (8)^2 + (10)^2 + (3)^2 + (5)^2 + (8)^2 + (9)^2 1] - 660.08$$

$$= [25 + 81 + 121 + 100 + 16 + 49 + 64 + 100 + 9 + 25 + 64 + 81] - 660.8$$

$$= 735 - 660.08$$

$\therefore$ TSS = 74.92

Sum of Squares Between Soya Bean (Columns)

$$SSB = \sum_j \frac{T_j^2}{n_j} - \frac{(GT)^2}{N}$$

$$= \frac{(12)^2}{3} + \frac{(21)^2}{3} + \frac{(27)^2}{3} + \frac{(29)^2}{3} - 660.08$$

$$= [48 + 147 + 243 + 280.33] - 660.08$$

$$= 718.33 - 660.08$$

$\therefore$ SSB = 58.25

Degree of freedom (r) = (K – 1)

= (4 – 1)

= 3

Mean sum of squares between the soya beans = $\frac{58.25}{3}$ = 19.42

Sum of Squares within Blocks (SSW)

$$SSW = TSS - SSB$$
$$= 74.92 - 58.25$$
$$= 16.67$$

Mean sum of squares within the blocks

$$= \frac{16.67}{12-4} = \frac{16.67}{8} = 2.08$$

ANOVA TABLE

Sources of variation	Sum of squares	Degrees of Freedom	Means squares
Between soya bean type	58.25	(k – 1) = (4 – 1) = 3	19.42
Within blocks	16.67	(n – k) = (12 – 4) = 8	2.08
Total		(n – 1) = (12 – 1) = 11	

$$\text{F-Ratio} = \frac{\text{Mean square between soya bean type}}{\text{Mean square within blocks}}$$

$$= \frac{19.42}{2.08} = 9.34$$

[Note: Assuming level of significance as 5%]

$\text{F-Ratio}_{(3, 8)}$, calculated = 9.34

F-Ratio from table for $V_1 = 3$ and $V_2 = 8$ at 5% level of significance = 4.07

Decision

The calculated value of F is more than the table value. Therefore we reject null hypothesis (H_0) which means that there is a significant difference between types of soya beans.

5.2.2 Inferences about Two Population Variances

Q9. Explain briefly about F-Distribution.

Ans : **(Imp.)**

The F-test is named in honour of the great statistician R.A. Fisher. The object of the F-test is to find out whether the two independent estimates of population variance differ significantly, or whether the two samples may be regarded as drawn from the normal populations having the same variance. For carrying out the test of significance, we calculate the ratio F. F is defined as :

$$F = \frac{S_1^2}{S_2^2}, \text{ where } S_1^2 = \frac{\left(X_1 - \bar{X}_1\right)^2}{n_1 - 1}$$

and $S_2^2 = \frac{\Sigma(X_2 - X_2)^2}{n_2 - 1}$

It should be noted that S^2 is always the larger estimate of variance, Le., $S_1^2 > S_2^2$

$$F = \frac{\text{Larger estimate of variance}}{\text{Smaller estimate of variance}}$$

$v_i = n_1 - 1$ and $v_2 = n_2 - 1$

v_1 = degrees of freedom for sample having larger variance.

v_2 = degrees of freedom for sample having smaller variance.

The calculated value of F is compared with the table value for v_1 and v_2 at 5% or 1% level of significance. If calculated value of F is greater than the table value then the F ratio is considered significant and the null hypothesis is rejected. On the other hand, if the calculated value of F is less than the table value the null hypothesis is accepted and it is inferred that both the samples have come from the population having same variance.

Since F test is based on the ratio of two variances, it is also known as the Variance Ratio Test. The ratio of two variances follows a distribution called the F distribution named after the famous statistician R.A. Fisher.

Assumptions in F-Test :

1. **Normality**

   The values in each group are normally distributed.

2. **Homogeneity**

   the variance within each group should be equal for all groups ($\sigma_1^2 = \sigma_2^2 = ... = \sigma_c^2$)

   This assumption is needed in order to combine or pool the variances within the groups into a single 'within groups' source of variation.

3. **Independence of error**

   It states that the error (variation of each value around its own group mean) should be independent for each value.

PROBLEMS

9. Two random samples were drawn from two normal populations and their values are :

A :	66	67	75	76	82	84	88	90	92		
B :	64	66	74	78	82	85	87	92	93	95	97

Test whether the two populations have the same variance at the 5% level of significance. (F = 36) at 5% level for $v_1 = 10$ and $v_2 = 8$.

Sol :

Let us take the hypothesis that the two populations have the same variance. Applying F-test

$$F = \frac{S_1^2}{S_2^2}$$

A $(X_1 - \bar{X}_1)$			B $(X_2 - \bar{X}_2)$		
X_1	x_1	x_1^2	X_2	x_2	x_2^2
66	–14	196	64	–19	361
67	–13	169	66	–17	289
75	–5	25	74	–9	81
76	–4	16	78	–5	25
82	+2	4	82	–1	1
84	+4	16	85	+2	4
88	+8	64	87	+4	16
90	+10	100	92	+9	81
92	+12	144	93	+10	100
			95	+12	144
			97	+14	196
$\Sigma X_1 = 720$	$\Sigma x_1 = 0$	$\Sigma x_1^2 = 734$	$\Sigma X_2 = 913$	$\Sigma x_2 = 0$	$\Sigma x_2^2 = 1298$

$$\bar{X}_1 = \frac{\Sigma X_1}{n_1} = \frac{720}{9} = 80; \quad \bar{X}_2 = \frac{\Sigma X_2}{n_2} = \frac{913}{11} = 83$$

$$S_1^2 = \frac{\Sigma x_1^2}{n_1 - 1} = \frac{734}{9-1} = 91.75$$

$$S_2^2 = \frac{\Sigma x_2^2}{n_2 - 1} = \frac{1298}{11-1} = 129.8$$

$$F = \frac{S_1^2}{S_2^2} = \frac{129.8}{91.75} = 1.415$$

For $v_1 = 10$ and $v_2 = 8$, $F_{0.05} = 3.36$

The calculated value of F is-less than the table value. The hypothesis is accepted. Hence it may be calculated that the two populations have the same variance.

10. In a sample of 8 observations, the sum of squared deviations of items from the mean was 84 4. In another sample of 10 observations, the value was found to be 102. 6. Test whether the difference is significant at 5% level.

You are given that at 5% level, critical value of F for v = 7 and V2 = 9 degrees of freedom is 3.29 and for $v_1 = 8$ and $v_2 = 10$ degrees of freedom, its value is 3.07

Sol :

Let us take hypothesis that the difference in the variances of the two samples is not significant. We are given :

$n_1 = 8, \Sigma\left(X_1 - \bar{X}_2\right)^2 = 84.4$

$n_2 = 10, \Sigma\left(X_2 - \bar{X}_2\right)^2 = 102.6$

$$F = \frac{S_1^2}{S_2^2}$$

$$S_1^2 = \frac{\Sigma\left(X_2 - \bar{X}_2\right)^2}{n_2 - 1} = \frac{84.4}{7} = 12.06$$

$$S_2^2 = \frac{\Sigma\left(X_2 - \bar{X}_2\right)^2}{n_2 - 1} = \frac{102.6}{9} = 11.4$$

$$F = \frac{12.06}{11.4} = 1.06$$

For $v_1 = 7$ and $v_2 = 9$ $F_{0.05} = 3.29$.

The calculated value of F is less than the table value. Hence we accept the hypothesis and conclude that the difference in the variance of two samples is not significant at 5% level:

11. Two samples are drawn from two normal population. From the following date test whether the two samples have the same variance at 5% level :

Sample 1 :	60	65	71	74	76	82	85	87		
Sample 2 :	61	66	67	85	78	63	85	86	88	91

Sol :

Let us take the hypothesis that the two populations have the same variance Applying F-test :

$$F = \frac{S_1^2}{S_2^2}$$

Sample 1 $(X_1 - \bar{X})_1$			Samples 2 $(X_2 - \bar{X}_2)$		
X_1	x_1	x_1^2	X_2	x_2	x_2^2
60	–15	225	61	-16	256
65	–10	100	66	-11	121
71	–4	16	67	-10	100
74	–1	1	85	+8	64
76	+1	1	78	+1	1
82	+7	49	63	-14	196
85	+10	100	85	+8	64
87	+12	144	86	+9	81
			88	+11	121
			91	+14	196
$\Sigma X_1 = 600$	$\Sigma x_1 = 0$	$\Sigma X_1^2 = 636$	$\Sigma X_2 = 770$	$\Sigma x_2 = 0$	$\Sigma x_2^2 = 1200$

$$\bar{X}_1 = \frac{600}{8} = 75;\ \bar{X}_2 = \frac{770}{10} = 77$$

$$S_1^2 = \frac{\Sigma x_1^2}{n_1 - 1} = \frac{636}{8-1} = \frac{636}{7} = 90.857$$

$$S_2^2 = \frac{\Sigma x_2^2}{n_2 - 1} = \frac{1200}{10-1} = \frac{1200}{9} = 133.333$$

$$F = \frac{133.333}{90.857} = 1.468$$

For $v_1 = 9$ and $v_2 = 7$, $F_{0.05} = 3.68$. The calculated value of F is less than the table value. The hypothesis holds good and hence we conclude that the two populations have the same variance.

5.3 CORRELATION

Q10. Define Correlation. Explain different types of Correlation.

Ans : **(Imp.)**

Meaning

Correlation refers to the relationship of two or more variables.

Correlation is a statistical analysis which measures and analyses the degree or extent to which two variables fluctuate with reference to each other.

The correlation expresses the relationship or interdependence of two sets of variables upon each other. One variable may be called the subject (independent) and the other rela (dependent).

Types Of Correlation

Correlation is classified into many types.

1. Positive and negative
2. Simple and multiple
3. Partial and total
4. Linear and non-linear

1. Positive and Negative Correlation

Positive and Negative correlation depend upon the direction of change of the variables. If two variables tend to move together in the same direction i.e. an increase in the value of one variable is accompanied by an increase in the value of the other variable; or a decrease in the value of one variable is accompanied by a decrease in the value of the other variable, then the correlation is called positive or direct correlation. Height and weight, rainfall and yield of crops, price and supply are examples of positive correlation.

If two variables, tend to move together in opposite directions so that an increase or decrease in the values of one variable is accompanied by a decrease or increase in the value of the other variable, then the correlation is called negative or inverse correlation.

2. Simple and Multiple correlation

When we study only two variables, the relationship is described as simple correlation; example are quantity of money and price level, demand and price, etc. But in a multiple correlation we study more than two variables simultaneously; example is the relationship of demand and supply of a commodity.

3. Partial and Total correlation

The study of two variables excluding some other variables is called partial correlation. For example, we study price and demand, eliminating the supply side. In total correlation, all the facts are taken into account.

4. Linear and non-linear correlation

If the ratio of change between two variables is uniform, then there will be linear correlation between them. Consider the following

A	2	7	12	17
B	3	9	15	21

We can see that the ratio of change between the variables is the same. If we plot these on the graph, we get a straight line.

In a curvilinear or non - linear correlation, the amount of change in one variable does not bear a constant ratio of the amount of change in the other variables. The graph of non linear or curvilinear relationship will be a curve.

Q11. Explain the Properties of Correlation.

Ans :

i) The value of correlation 'r' varies between [–1, +1]. This indicates that the r values does not exceed unity.

ii) Sign of the correlation sign of the Covariance.

iii) If r = –1 variables and perfectly negatively correlated.

iv) If r = + 1 variables are perfectly positively correlated.

If r = 0 variables are not correlated in a linear fashion. There may be non-linear relationship between variables.

Correlation coefficient is independent of change of scale and shifting of origin. In other words, Shifting the origin and change the scale do not have any effect on the value of correlation.

Q12. State the Limits for Coefficient of Correlation.

Ans :

Limits for Coefficient of Correlation for (x, y)

The value of the coefficient of correlation should lie between +1 and –1. If r = +1, the correlation is perfect and positive and if r = –1 the correlation is perfect and negative. When r = 0, it means that there is no relationship between the two variables.

Hence, $-1 < r(x, y) \leq 1$

Note

The correlation coefficient describes not only the magnitude of correlation but also its direction. Thus, +1 would mean that correlation is positive and the magnitude of correlation is 1.

Similarly -1 means correlation is negative and the magnitude of correlation is again 1.

S.No.	Degree of Correlation	The Value of r(x,y) (Positive and Negative)
1.	Perfect correlation	Exactly 1
2.	Very high degree of correlation	0.90 and above but less than 1
3.	Fairly high degree of correlation	0.75 and above but less than 0.90
4.	Moderate degree of correlation	0.50 and above but less than 0.75
5.	Low degree of correlation	0.25 and above but less than 0.50
6.	Very high degree of correlation	Below 0.25
7.	Absence of correlation	Equal to 0

Q13. What is Karl Pearson's Coefficient of Correlation? Explain properties of Coefficient of Correlation.

Ans : **(Imp.)**

Karl Parson's Coefficient of Correlation is arrived at with the help of a statistical formula that takes into account the mean and standard deviation of the two variables, the number of such observations and the covariance between them. Since Karl Pearson's coefficient of correlation is a number, it can describe the strength of the correlation in greater detail and more objectively. A value of –1 signifies "absolute" negative correlation, a value between – 1 and – 0.5 signifies strong negative correlation, a value between – 0.5 and – 0.25 signifies moderate negative correlation and a value between – 0.25 and 0 signifies weak negative correlation. Similarly, a value of +1 signifies "absolute" positive correlation, a value between +1 and + 0.5 signifies strong positive correlation, a value between + 0.5 and + 0.25 signifies moderate positive correlation and a value between + 0.25 and 0 signifies weak positive correlation

Properties of Karl Pearson's Coefficient of Correlation

1. It is based on Arithmetic Mean and Standard Deviation.
2. It lies between –1
3. It measures both direction as well as degree of change. If r is less than 0, there is negative correlation, which means the direction of change of the two variables will be opposite. If r is more than 0, there is positive correlation, which means the direction of change of the two, variables will be same. Higher the value of r, greater is the degree of correlation. Hence, Karl Pearson's coefficient of correlation is said to be the ideal measure of correlation.
4. It is independent of change in scale. In other words, if a constant amount is added/ subtracted from all values of a variable, the value of r does not change.
5. It is independent of change in origin. Thus, if a constant amount is multiplied with or divides all values of a variable, *Y* does not change.

6. It is independent of direction. In other words, Correlation of X and Y is same as Correlation of Y and X.
7. It is a pure number without any units. In other words, it is independent of the unit of measurement of the 2 variables.
8. It takes into account all items of the variable(s).
9. It does not prove causation but is simply a measure of co-variation.
10. Correlation coefficient of two variables X and Y is the Geometric Mean of two regression coefficients, regression coefficient of X on Y and regression coefficient of Y on X. Symbolically,

    r = Square root of (b_{xy} * b_{yx})
11. Correlation coefficient can be calculated between two unrelated variables and such a number can be misleading. Such correlation is called accidental correlation, spurious correlation or non sense correlation.

Q14. Explain the methods of Coefficient of Correlation.

Ans :

i) Direct Method when deviations are taken from actual mean

$$r = \frac{\Sigma xy}{N\sigma_x\alpha_y}$$

However, this formula is transformed in the following form

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \times \Sigma y^2}}$$

Where

$$x = X - \bar{X} \text{ and } y = Y - \bar{Y}$$

Steps :

1. Find the means of the two series ($\bar{X}$, $\bar{Y}$)
2. Take the deviations of X series from the mean of X and denote these deviations as x.
3. Square these deviations and obtain the total. Denote it as Σx^2.
4. Take the deviations of Y series from the Mean of Y and denote these deviations as y.
5. Square these deviations, obtain the total and denote it as Σy^2.
6. Multiply the deviations of X and Y series, obtain the total and denote it Σxy.
7. Substitute the above values in the formula.

ii) Short-Cut Method

When deviations are taken from assumed mean.

When actual mean is in fraction, then the above formula becomes tedious. In such cases, assumed mean is used for calculating correlation. The formula is.

$$r = \frac{\Sigma dxdy - \frac{\Sigma dx.\Sigma dy}{N}}{\sqrt{\Sigma dx^2 - \frac{(\Sigma dx)^2}{N}}\sqrt{\Sigma dy^2 - \frac{(\Sigma dy)^2}{N}}}$$

Where

$\Sigma dxdy$ = Sum of the product of the deviations of X and Y series from their assumed means.

Σdx^2 = Sum of the squares of the deviations of X series from an assumed mean.

Σdy^2 = Sum of the squares of the deviations of Y series from an assumed mean.

Σdx = Sum of the deviations of X series from an assumed mean.

Σdy = Sum of the deviations of Y Series from an assumed mean.

N = No. of Pairs of observations.

The values of coefficient of correlation as obtained by above formulae will always lie between ± 1. When there is perfect positive correlation its value is +1 and when there is perfect negative correlation, its value is –1. When r = 0 means that there is no relationship between the two variables. We normally get values which lie between +1 and –1.

Probable Error of the Coefficient of correlation and its interpretation.

The probable error of the coefficient of correlation helps in interpretation. The probable error of the coefficient of correlation is obtained as follows:

$$\text{P.E. of } r = 0.6745 \frac{1-r^2}{\sqrt{N}}$$

Where

r = Coefficient of correlation;

N = Number of pairs of observations.

If the probable error is added to and subtracted from the coefficient of correlation, it would give two such limits within which we can reasonably expect the value of coefficient of correlation to vary.

Symbolically P (rho) = r ± P.E.

Where 'P' denotes the correlation in the population. Suppose, the Coefficient of correlation for a pair of 10 observations is 0.8 and its P.E. is 0.05. The limits of the correlation in the population would be r ± P.E. i.e. 0.8 ± 0-.05 or 0.75 – 0.85.

If the value of r is less than the probable error then r is not at all significant, i.e. there is no evidence of correlation. If the value of r is more than six times the probable error, it is significant. Hence it can be said that r is significant, when

$$r > 6 \text{ P.E. or } \frac{r}{\text{P.E}} > 6$$

PROBLEMS

12. Find the coefficient of correlation between x and y

X :	65	66	67	67	68	69	70	72
Y :	67	68	65	68	72	72	69	71

Sol :

X	Y	$x = X - \bar{X}$	x^2	$y = Y - \bar{Y}$	y^2	xy
65	67	–3	9	–2	4	6
66	68	–2	4	–1	1	2
67	65	–1	1	–4	16	4
67	68	–1	1	–1	1	1
68	72	0	0	0	9	0
69	72	1	1	3	9	3
70	69	2	4	0	0	0
72	71	4	16	2	4	8
544	552	$\Sigma x = 0$	$\Sigma x^2 = 36$	$\Sigma y = 0$	$\Sigma y^2 = 44$	$\Sigma xy = 24$

$$\bar{X} = \frac{\Sigma x}{n} = \frac{544}{8} = 68,$$

$$\bar{Y} = \frac{\Sigma y}{n} = \frac{552}{8} = 69$$

As per Karl Perason, coefficient of correlation

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \times \Sigma y^2}} = \frac{24}{\sqrt{36 \times 44}} = +0.603$$

13. Find if there is any significant correlation between the heights and weights given below.

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

]

Sol :

Computation of coefficient of correlation

Height in inches x	Deviation from Mean (60) $X = x - \bar{X}$	Square of deviations X^2	Weight in lbs y	Deviations from Mean $Y = y - \bar{y}$	Square of deviations Y^2	Product of deviations of X and Y series (XY)
57	– 3	9	113	– 7	49	21
59	– 1	1	117	– 3	9	3
62	2	4	126	6	36	12
63	3	9	126	6	36	18
64	4	16	130	10	100	40
65	5	25	129	9	81	45
55	– 5	25	111	– 9	81	45
58	– 2	4	116	– 4	16	8
57	– 3	9	112	– 8	64	24
540	0	102	1080	0	472	216

Coefficient of correlation $r = \frac{\Sigma XY}{\sqrt{\Sigma X^2 \times \Sigma Y^2}}$

$\therefore r = \frac{216}{\sqrt{102 \times 471}} = 0.98$

14. Calculate the coefficient of correlation for the following age of husbands and wifes.

Husband's Age (in years)	23	27	28	29	30	31	33	35	36	39
Wife's Age (in years)	18	22	23	24	25	26	28	29	30	32

Sol :

Age of Husband	X–A		Age of Y–A Wife			
X	A = 30 dx	dx^2	Y	A = 25 dy	dy^2	dxdy
23	–7	49	18	–7	49	+49
27	–3	9	22	–3	9	+9
28	–2	4	23	–2	4	+4
29	–1	1	24	–1	1	+1
30	0	0	25	0	0	+0
31	+1	1	26	+1	1	+1
33	+3	9	28	+3	9	+9
35	+5	25	29	+4	16	+20
36	+6	36	30	+5	25	+30
39	+9	81	32	+7	49	+63
	$\Sigma dx = +11$	$\Sigma dx^2 = 215$		$\Sigma dy = +7$	$\Sigma dy^2 = 163$	$\Sigma dxdy = 186$

$$r = \frac{\Sigma dxdy - \frac{\Sigma dx.\Sigma dy}{N}}{\sqrt{\Sigma dx^2 - \frac{(\Sigma dx)^2}{N}}\sqrt{\Sigma dy^2 - \frac{(\Sigma dy)^2}{N}}}$$

$\Sigma dxdy = 186$, $\Sigma dx = 11$,
$\Sigma dy = +7$, $\Sigma dx^2 = 215$,
$\Sigma dy^2 = 163$, $N = 10$

$$\frac{185 - \frac{11 \times 7}{10}}{\sqrt{215 - \frac{(11)^2}{10}}\sqrt{163 - \frac{(7)^2}{10}}}$$

$$\frac{186 - 7.7}{\sqrt{215 - 12.1}\sqrt{163 - 4.9}} \quad \frac{178.3}{14.244 \times 12.574}$$

$$= \frac{178.3}{179.104}$$

$$= 0.966$$

There is a very high degree of positive correlation between the ages of husbands and wives.

5.4 REGRESSION

Q15. Define Regression. State the uses of Regression.

Ans : **(Imp.)**

The dictionary meaning of the term 'regression' is the act of the returning or going back. The term 'regression' was first used by Sir Francis Galton in 1877 while studying the relationship between the heights of father and sons. Regression analysis is a technique used for the modeling and analysis of numerical data consisting of values of a dependent variable (response variable) and of one or more independent variables.

1. **Dependant Variable** is the single variable being explained/ predicted by the regression model (response variable).
2. **Independent Variable** is the explanatory variable(s) used to predict the dependant variable (Predictor variable).

Definitions

(i) According to "Regression is the measure of the average relationship between two or more variables in terms of the original units of data."

(ii) According to Blair, "Regression is the measure of the average relationship between two or more variable in terms of the original units of the data."

(iii) According to Taro Yamane, "One of the most frequently used techniques in economics and business research, to find a relation between two or more variable that are related causally, is regression analysis."

Uses

1. It is used to estimate the relation between two economic variables like Income and Expenditure.
2. It is a highly valuble tool in Economics and Business.
3. It is widey used for prediction purpose.
4. We can calculate coefficient of correlation and coefficient of determination with the help of the regression coefficient.
5. It is useful in statistical estimation of demand curves, supply curves, production function, cost function and consumption function etc.

Q16. What are the objectives of Regression Analysis?

Ans :

1. The first objective of regression analysis is to provide estimates of values of the dependent variable from values of independent variable. This is done with the help of the regression line. The regression line describes the average relationship existing between X and Y variables, more precisely, it is a line which displays mean values of Y for given values of X.
2. The second objective of regression analysis is to obtain a measure of the error involved in using the regression line as a basis for estimation. For this purpose standard error of estimate is obtained. This helps in understanding the correlation existing between X and Y.
3. In general, we can model the expected value of y as an n^{th} order polynomial, yielding the general polynomial regression model

$$Y = a_0 + a_1x + a_2x^2 + a_2x^2 + a_3x^3 + \ldots + a_nx^n + \varepsilon$$

Conveniently, these models are all linear from the point of view of estimation, since the regression function is linear in terms of the unknown parameters ao,ai,.... Therefore, for least squares analysis, the computational and inferential problems of polynomial regression can be completely addressed using the techniques of multiple regressions. This is done by treating x, x^2, ... as being distinct independent variables in a multiple regression model.

Q17. What are the assumptions of Regression Analysis.

Ans :

The following assumptions are made while making use of the regression technique:

1. There exists an actual relationship between the dependent and independent variables.
2. The regression analysis is used to estimate the values within the range for which it is valid and not for the values outside its range.
3. The relationship that existed between the dependent and independent variables remains the same till the regression equation is calculated.
4. The dependent variable takes any random value but the values of the independent variables are fixed quantities without error and are chosen by the analyst or the user.
5. In regression, we have only one .dependent variable in our estimating equation. However, we can use more than one independent variable.

5.4.1 Limitations

Q18 What are the limitations of Regression Analysis?

Ans : **(Imp.)**

1. It assumes a linear relationship between two variables which need not be the case always.
2. It assumes a static relationship between the two variables over a period of time. However, relationships between variables can change with a change in other factors. For example, the change in demand for a given change in price can be estimated using regression. However, the impact of price on demand will be different when a family or a nation is poor and when such a family or nation has abundance of wealth or resources.
3. Regression analysis provides meaningful insights only up to a certain limit. For example, increasing production results in a decrease in marginal cost. However, beyond a certain point, increase in production can result in the costs going up.

Q19. Explain different types of Regression.

Ans :

The various types of Regression are as follows:

1. Simple Regression

In statistics, simple regression is the least squares estimator of a linear regression model with a single predictor variable. In other words, simple linear regression fits a straight line through the set of n points in such a way that makes the sum of squared residuals of the model (that is, vertical distances between the points of the data set and the fitted line) as small as possible.

2. Multiple Regression

Multiple regression analysis represents a logical extension of two-variable regression analysis. Instead of a single independent variable, two or more independent variables are used to estimate the values of a dependent variable. However, the fundamental concepts in the analysis remain the same.

For example, a college admissions officer wishing to predict the future grades of college applicants might use three variables (High School GPA, SAT and Quality of letters of recommendation) to predict college GPA. The applicants with the highest predicted college GPA would be admitted. The prediction method would be developed based on students already attending college and then used on subsequent classes. Predicted scores from multiple regression are linear combinations of the predictor variables. Therefore, the general form of a prediction equation from multiple regression is:

$$Y' = b, X, + b_2 X_2 + ... + b_k X_k + A$$

where Y' is the predicted score, *X,* is the score on the first predictor variable, X_2 is the score on the second, etc. The Y intercept is A. The regression coefficients (b_1, b_2, etc.) are analogous to the slope in simple regression.

3. Curvilinear Regression

The analysis of the linear regression model can be extended in a straightforward way to cover situations in which the dependent variable is affected by several controlled variables or in which it is affected non-linearly by one controlled variable.

For example, suppose that there are three controlled variables, x,, x_2 and x_3. A linear regression equation is of the form,

$$y = a_0 + a_1 x + a_2 x^2 + a_3 x^3$$

4. Polynomial Regression

Suppose that the dependent variable is a polynomial function of a single controlled variable. For example, in cubic regression, the regression equation is given by,

$$y = ao + ai\, x + a_2 x^2 + a3X^3$$

This type of regression can be approached in the same way as multiple regressions. In the case of cubic regression we can substitute x, = x, $x_2 = x^:$ and x, = x^3. The least squares estimates of a_0, a,, a_2 and a_3 can then be obtained.

If the observations are taken in such a way that there are an equal number of observations on y at a series of equally spaced values of x, then it is computationally more efficient to use the method of orthogonal polynomials.

5.4.2 Estimation using Regression Line

Q20. State the various ways to Estimate the regression line.

Ans :

Regression Equations

The regression equations express the regression lines. As there are two regression lines, so there are two regression equations. The regression equation X on Y describes the variation in the values of X for the given changes in Y, and used for estimating the value of X for the given value of Y, Similarly, the regression equation Y on X describes the variation in the values of Y for the given changes in X, and is used for estimating the value of Y for the given value of X.

1. Regression Equation of Y on X : With a little algebra and differential calculus it can be shown that the following two equations, if solved simultaneously, will yield values of the parameters a and b such that the least squares requirement is fulfilled:

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

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

These equations are usually called the normal equations. In the equations ΣX, ΣXY, ΣX^2 indicate totals which are computed from the observed pairs of values of two variables X and Y to which the least squares estimating line is to be fitted and N is the number of observed pairs of values.

2. Regression Equation of X on Y : The regression equation of X on Y is expressed as follows:

$$X = a + bY$$

To determine the values of a and b, the following two normal equations are to be solved simultaneously:

$$\Sigma Y = Na + b\Sigma Y \qquad \Sigma XY = a\Sigma Y + b\Sigma Y^2$$

PROBLEMS

15. From the following data obtain the two regression equations :

X :	6	2	10	4	8
Y :	9	11	5	8	7

Sol :

X	Y	XY	X^2	Y^2
6	9	54	36	81
2	11	22	4	121
10	5	50	100	25
4	8	32	16	64
8	7	56	64	49
$\Sigma X = 30$	$\Sigma Y = 40$	$\Sigma XY = 214$	$\Sigma X^2 = 220$	$\Sigma Y^2 = 340$

Table : Obtaining Regression Equations

Regression equation of Y on X, $Y = a + bX$

To determine the values of a and b the following two normal equations are to be solved.

$$\Sigma Y = Na + b\Sigma Y$$

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

Substituting the values

$$40 = 5a + 30b \quad \text{... (1)}$$

$$214 = 30a + 220b \quad \text{... (2)}$$

Multiplying equation (1) by 6, $240 = 30a + 180b$ ... (3)

$$214 = 30a + 220b \quad \text{.... (4)}$$

Deducing equation (4) from (3) $-40b = 26$ or $b = -0.65$

Substituting the value of b in equation (1)

$$40 = 5a + 30(-0.65)$$

or $5a = 40 + 19.5 = 59.5$ or $a = 11.9$

Putting the values of a and b in the equation

Regression equation of Y on X

$$Y = 11.9 - 0.65X$$

Regression equation of X on Y $X = a + bY$ and the two normal equations are:

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

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

Substituting the values

$30 = 5a + 40b$ ... (5)

$214 = 40a + 340b$ ... (6)

Multiplying equation (5) by 8:

$240 = 40a + 320b$ ... (7)

$214 = 40a + 340b$ ... (8)

From equation (7) and (8)

$-20b = 26$ or $b = -1.3$

Substituting the value of b in equation (5);

$30 = 5a + 40(-1.3)$

$5a = 30 + 52 = 82$

$\therefore \quad a = 16.4$

Putting the value of a and b in the equation.

The regression equation of X of Y :

$X = 16.4 - 1.3\,Y.$

16. From the following data, calculate the regression equations taking deviation of items from the mean of X and Y series.

X	6	2	10	4	8
Y	9	11	5	8	7

Sol :

$X - \bar{X} = bxy\,(Y - \bar{Y})$

$\bar{X} = \frac{\Sigma X}{N} = \frac{30}{5} = 6$

$\bar{Y} = \frac{\Sigma Y}{N} = \frac{40}{5} = 8$

$\therefore \quad \bar{X} = 6, \ \bar{Y} = 8$

X	Y	$X - \bar{X}$ (x)	$Y - \bar{Y}$ (y)	$(X - \bar{X})^2 = x^2$	$(Y - \bar{Y})^2 = y^2$	xy
6	9	0	1	0	1	0
2	11	–4	3	16	9	–12
10	5	4	–3	16	9	–12
4	8	–2	0	4	0	0
8	7	2	–1	4	1	–2
$\Sigma X = 30$	$\Sigma Y = 40$	$\Sigma x = 0$	$\Sigma y = 0$	$\Sigma x^2 = 40$	$\Sigma y^2 = 20$	$\Sigma xy = 26$

Number of Pairs N = 5

Regression Equation of X on Y

Regression Coefficients

$$b_{xy} = \frac{\Sigma xy}{\Sigma y^2} = \frac{-26}{20} = -1.3$$

$$\therefore \quad X - \bar{X} = b_{xy}(Y - \bar{Y})$$

$$X - 6 = 1.3(Y - 8)$$

$$X - 6 = 1.3Y + 10.4$$

$$X = -1.3Y + 10.4 + 6$$

$$X = 1.3Y + 16.4$$

(or)

$$X = 16.4 - 1.3Y$$

Regression Equation of Y on X

$$(Y - \bar{Y}) = b_{yx}(X - \bar{X})$$

Regression Coefficient

$$b_{yx} = \frac{\Sigma x^2}{\Sigma y^2} = \frac{-26}{40} = -0.65$$

$$\therefore \quad Y - \bar{Y} = b_{yx}(X - \bar{X})$$

$$Y - 8 = -0.65(X - 6)$$

$$Y - 8 = -0.65x + 3.9$$

$$Y = -0.65x + 3.9 + 8$$

$$Y = -0.65x + 11.9$$

or

$$Y = 11.9 - 0.65x$$

17. Determine the equation of a straightline which best fits the data.

X :	10	12	13	16	17	20	25
Y :	10	22	24	27	29	33	37

Sol :

Let the required traight lines $y = a + bX$

The two normal equations are $\Sigma Y = b\Sigma X + Na$

$$\Sigma XY = b\Sigma X^2 + a\Sigma x$$

X	X^2	Y	XY
10	100	10	100
12	144	22	264
13	169	24	312
16	256	27	432
17	289	29	493
20	400	33	660
25	625	37	925
$\Sigma X = 113$	$\Sigma X^2 = 1938$	$\Sigma Y = 182$	$\Sigma XY = 3186$

Substituting the values

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

$\Sigma Y = 182; \Sigma X = 113; N = 7$

$113b + 7a = 182$ .....(1)

$\Sigma XY = 3186 ; \Sigma X^2 = 1983; \Sigma X = 113$

$1983b + 113a = 3186$ .....(2)

Multiplying (1) by 113;

$12769b + 791a = 20566$ .....(3)

Multiplying (2) by 7;

$13881b + 791a = 22302$ ..... (4)

Subtracting (4) from (3); $b = \frac{1736}{1112} = 1.56 \Rightarrow a = 0.82$

The equation of straight line is

$Y = a + bX$

$a = 0.82; b = 1.56$

$Y = 0.82 + 1.56 X$

$\therefore$ The equation of the required straight line is $Y = 0.82 + 1.56 X$

This is called regression equation of Y on X.

5.4.3 Making inferences about population parameters

Q21. Write about various inferences about population parameters.

Ans :

1. Regular maintenance that does not depend on the age of the truck: tune-ups, oil changes, and lubrication. This expense is captured in the intercept term A.

$$Y = A + BX$$

2. Expenses for repairs due to aging: relining brakes, engine and transmission overhauls, and painting. Such expenses tend to increase with age of the truck, and they are captured in the BX term of the population regression line

$$Y = A + BX.$$

The individual data points will satisfy the formula

$$Y = A + BX + e$$

where e is a random disturbance from the population regression line.

On the average, e equals zero because disturbances above the population regression line are canceled out by disturbances below the line. We can denote the standard deviation of these individual disturbances by σ_e. The standard error of estimate s_e, then, is an estimate of s_e, the standard deviation of the disturbance.

Slope of the Population Regression Line

The regression line is derived from a sample and not from the entire population. As a result, we cannot expect the true regression equation. Y = A + BX (the one for the entire population), to be exactly the same as the equation estimated from the sample observation, (or) $\hat{Y} = a + bX$. Even so, we can use the value of b, the slope we calculate from a sample, to test hypotheses about he value of B, the slope of the regression line for the entire population.

Suppose that over an extended past period of time, the slope of the relationship between X and Y was 2.1. To test whether this is still the case, we could define the hypotheses as,

$H_0 : B = 2.1 \leftarrow$ Null Hypothesis

$H_1 : B \neq 2.1 \leftarrow$ Alternative Hypothesis

In effect, then, we are testing to learn whether current data indicate that B has changed from its historical value of 2.1.

To find the test statistic for B, it is necessary first to find the standard error of the regression coefficient. Here, the regression coefficient we are working with is b, so the standard error of this coefficient is denoted s_b. It presents the mathematical formula for s_b:

$$s_b = \frac{s_e}{\sqrt{\Sigma X^2 - nX^2}}$$

where

- s_b = standard error of the regression coefficient
- s_e = standard error of estimate
- X = values of the independent variable
- $\bar{X}$ = mean of the values of the independent variable
- n = number of data points

Once we have calculated s_b, we can use equation to standardize the slope of our fitted regression equation:

Standardized Value of b

$$t = \frac{b - B_{H_0}}{s_b}$$

where

- b = slope of fitted regression
- B_{H_0} = actual slope hypothesized for the population
- s_b = standard error of the regression coefficient.

Because the test will be based on the t distribution with n – 2 degrees of freedom, we use t to denote the standardized statistic.

5.5 Errors and Caveats in Regression and Correlation Analysis

Q22. Explain briefly about errors and caveats in Regression and Correlation analysis.

Ans :

Standard errors of the regression and correlation analysis The standard error of the model (denoted again by s) is usually referred to as the standard error of the regression (or sometimes

the "standard error of the estimate") in this context, and it is equal to the square root of {the sum of squared errors divided by n–2}, or equivalently, the standard deviation of the errors multiplied by the square root of (n-1)/(n-2), where the latter factor is a number slightly larger than 1:

$$s = \sqrt{\frac{1}{n-2}\Sigma_{t=1}^{n} e_t^2}$$

= STDEV.S(errors) × SQRT((n – 1)/n – 2))

The sum of squared errors is divided by n – 2 in this calculation rather than n – 1 because an additional degree of freedom for error has been used up by estimating two parameters (a slope and an intercept) rather than only one (the mean) in fitting the model to the data. The standard error of the regression is an unbiased estimate of the standard deviation of the noise in the data, i.e., the variations in Y that are not explained by the model.

Each of the two model parameters, the slope and intercept, has its own standard error, which is the estimated standard deviation of the error in estimating it. (In general, the term "standard error" means "standard deviation of the error" in whatever is being estimated). The standard error of the intercept is

$$SEb_a = \frac{s}{\sqrt{n}} \times \sqrt{1+\frac{(Average(X))^2}{VARP.(X)}}$$

which looks exactly like the formula for the standard error of the mean in the mean model, except for the additional term of (AVERAGE(X))²/ VAR.P(X) under the square root sign. This term reflects the additional uncertainty about the value of the intercept that exists in situations where the center of mass of the independent variable is far from zero (in relative terms), in which case the intercept is determined by extrapolation far outside the data range. The standard error of the slope coefficient is given by:

$$SE_{b_1} = \frac{s}{\sqrt{n}} \times \frac{1}{STDEV.P(X)}$$

which also looks very similar, except for the factor of STDEV.P(X) in the denominator. Note that s is measured in units of Y and STDEV.P(X) is measured in units of X, so SE_{b1} is measured (necessarily) in "units of Y per unit of X", the same as b_1 itself. The terms in these equations that involve the variance or standard deviation of X merely serve to scale the units of the coefficients and standard errors in an appropriate way.

The standard errors of the coefficients are directly proportional to the standard error of the regression and inversely proportional to the square root of the sample size. This means that noise in the data (whose intensity if measured by s) affects the errors in all the coefficient estimates in exactly the same way, and it also means that 4 times as much data will tend to reduce the standard errors of the all coefficients by approximately a factor of 2, assuming the data is really all generated from the same model, and a really huge of amount of data will reduce them to zero.

Key Caveats with Correlations

There are three key caveats that must be recognized with regard to correlation.

1. It is impossible to prove causal relationships with correlation. However, the strength of the evidence for such a relationship can be evaluated by examining and eliminating important alternate explanations for the correlation seen.
2. Outliers can substantially inflate or deflate the correlation.
3. Correlation describes the strength and direction of the linear association between variables. It does not describe non-linear relationships

Correlation and Causation

It is often tempting to suggest that, when the correlation is statistically significant, the change in one variable causes the change in the other variable. However, outside of randomized experiments, there are numerous other possible reasons that might underlie the correlation.

1. Check for the possibility that the response might be directly affecting the explanatory variable (rather than the other way

around). For example, you might suspect that the number of times children wash their hands might be causally related to the number of cases of the common cold amongst the children at a pre-school. Check whether changes in the explanatory variable contribute, along with other variables, to changes in the response. For example, the amount of dry brush in a forest does not cause a forest fire; but it will contribute to it if a fire is ignited.

2. Check for confounders or common causes that may affect both the explanatory and response variables. For example, there is a moderate association between whether a baby is breast-fed or bottle-fed and the number of incidences of gastroenteritis recorded on medical charts.

3. Check whether both variables may have changed together over time or space. For example, data on the number of cases of internet fraud and on the amount spent on election campaigns in the United States taken over the last 30 years would have a strong association merely because they have both increased over time.

4. Check whether the association between the variables might be just a matter of coincidence. This is where a check for the degree of statistical significance would be important.

5.6 MULTIPLE REGRESSION

Q23. Define Multiple Regression. State the assumptions of Multiple Regression.

Ans :

Multiple regression generally explains the relationship between multiple independent or predictor variables and one dependent or criterion variable. A dependent variable is modeled as a function of several independent variables with corresponding coefficients, along with the constant term. Multiple regression requires two or more predictor variables, and this is why it is called multiple regression.

The multiple regression equation explained above takes the following form:

$$y = b_1x_1 + b_2x_2 + \ldots + b_nx_n + c$$

Where, bi's (i=1,2…n) are the regression coefficients, which represent the value at which the criterion variable changes when the predictor variable changes.

Assumptions

- There should be proper specification of the model in multiple regression. This means that only relevant variables must be included in the model and the model should be reliable.
- Linearity must be assumed; the model should be linear in nature.
- Normality must be assumed in multiple regression. This means that in multiple regression, variables must have normal distribution.
- Homoscedasticity must be assumed; the variance is constant across all levels of the predicted variable.

5.6.1 Finding multiple regression equations

Q24. How to find multiple regression equations.

Ans : **(Imp.)**

In simple regression, X is the symbol used for the values of the independent variable. In multiple regression, we have more than one independent variable. So we shall continue to use X, but we shall add a subscript (for example X_1, X_2) to distinguish between the independent variables we are using.

Example

We will let X_1 represent the number of field-audit labot hours and X_2 represent the number of computer hours. The dependent variable, Y, will be the actual unpaid taxes discovered.

In simple regression $\hat{Y} = a + bX$ describes the relationship between the two variables X and Y.

In multiple regression, we must extend that equation, adding one term for each new variable. In symbolic form, equation is the formula we can use when we have two independent variables:

Estimating Equation Describing Relationship Among Three Variables

$$\hat{Y} = a + b_1X_1 + b_2X_2 \quad \text{... (1)}$$

where

- $\hat{Y}$ = estimated value corresponding to the dependent variable.
- a = Y-intercept
- X_1 and X_2 = Values of the two independent variables.
- b_1 and b_2 = slopes associated with X_1 and X_2, respectively.

It can visualize the simple estimating equation as a line on a graph: similarly, we can picture a two-variable multiple regression equation as a plane, such as the one shown in Figure. Here we have a three dimensional shape that possesses depth, length, and width. To get an intuitive feel for this three-dimensional shape, visualize the intersection of the axes, Y, X_1, and X_2 as one corner of a room.

It is a graph of 10 sample points and the plane about which these points seem to cluster. Some points lie above the plane and some fall below It just as points lie above and below the simple regression line.

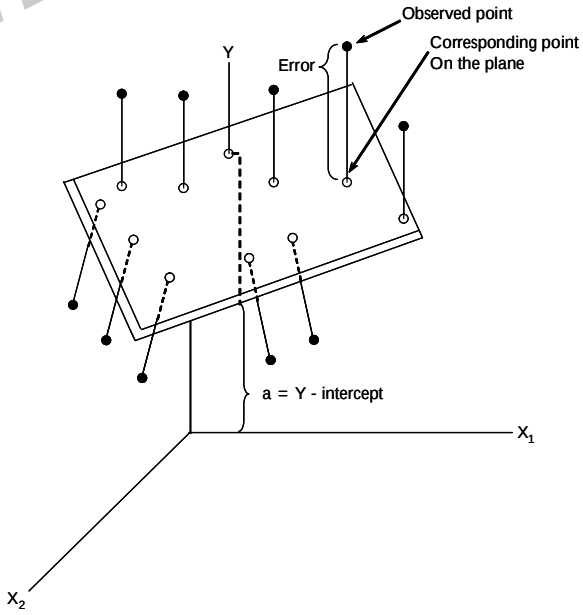
The problem is to decided which of the possible planes that we could draw will be the best fit. To do this, we shall again use the lest-squares criterion and locate the plane that minimizes the sum of the squares of the errors, that is, the distances from the points around the plane to the corresponding points on the plane. We use our data and the following three equations (which statisticians call the "normal equations") to determine the values of the numerical constant, a, b_1, and b_2.

Normal Equations

$$\Sigma Y = na + b_1\Sigma X_1 + b_2\Sigma X_2 \quad \text{... (2)}$$

$$\Sigma X_1Y = a\Sigma X_1 + b_1\Sigma X_1^2 + b_2\Sigma X_1X_2 \quad \text{... (3)}$$

$$\Sigma X_2Y = a\Sigma X_2 + b_1\Sigma X_1X_2 + b_2\Sigma X_2^2 \quad \text{... (4)}$$



Example :

Month :	January	February	March	April	May	June	July	August	September	October
X_1										
Field - Audit Labour Hours (00s Omitted)	45	42	44	45	43	46	44	45	44	43
X_2 Computer Hours (00s Omitted)	16	14	15	13	13	14	16	16	15	15
X_3 Actual Unpaid Taxes Discovered (millons of dollars)	29	24	27	25	26	28	30	28	28	27

Sol :

Y (1)	X_1 (2)	X_3 (3)	X_1Y (2)×(1)	X_2Y (3)×(1)	X_1X_2 (2)×(3)	X_1^2 $(2)^2$	X_2^2 $(3)^2$	Y^2 $(1)^2$
29	45	16	1,305	464	720	2,025	256	841
24	42	14	1,008	336	588	1,764	196	576
27	44	15	1,188	405	660	1,936	225	729
25	45	13	1,125	325	585	2,025	169	625
26	43	13	1,118	338	559	1,849	169	676
28	46	14	1,288	392	644	2,116	196	784
30	44	16	1,320	480	704	1,936	256	900
28	45	16	1,260	448	720	2,025	256	784
28	44	15	1,232	420	660	1,936	225	784
27	43	15	1.161	405	645	1,849	225	729
272	441	147	12.005	4,013	6,485	19,461	2,173	7,428
↑	↑	↑	↑	↑	↑	↑	↑	↑
ΣY	ΣX_1	ΣX_2	ΣX_1Y	ΣX_2Y	ΣX_1X_2	ΣX_1^2	ΣX_2^2	ΣY^2

$\bar{Y} = 27.2$

$\bar{X}_1 = 44.1$

$\bar{X}_2 = 14.7$

$$272 = 10a + 441b_1 + 147b_2 \quad .....(1)$$

$$12{,}005 = 441a + 19{,}461b_1 + 6{,}485b_2 \quad .....(2)$$

$$4{,}013 = 147a + 6{,}485b_1 + 2{,}173b_2 \quad .....(3)$$

When we solve these three equations simultaneously, we get

$$a = -13.828$$

$$b_1 = 0.564$$

$$b_2 = 1.099$$

5.6.2 Making Inferences about Population Parameters

Q25. Write about Inferences about individual slope Bi for population parameters.

Ans : **(Imp.)**

The regression plane is derived from a sample and not from the entire population. As a result, we cannot expect the true regression equation, $Y = A + B_1X_1 + B_2X_2 + ... + B_kX_k$ (the one for the entire population), to be exactly the same as the equation estimated from the sample observations, $\hat{Y} = a + b_1x_1 + b_2X_2 + .. + b_kX_k$. Even so, we can use the value of bi one of the slopes we calculate from a sample, to test hypotheses about the value of Bi, one of the slopes of the regression plane for the entire population.

To understand this process, return to the problem that related unpaid taxes discovered to field-audit labor hours, computer hours, and rewards to informants.

The first step is to find some value for B1 to compare with b1 = 0.597.

Suppose that over an extended past period of time, the slope of the relationship between Y and X1 was 0.400. To test if this were still the case, we could define the hypotheses as

H0: $B1 = 0.400 \rightarrow$ Null hypothesis

H1: $B1 \neq 0.400 \rightarrow$ Alternative hypothesis

In effect, then, we are testing to learn whether current data indicate that *B*1 has changed from its historical value of 0.400.

To find the test statistic for B1, it is necessary first to find the standard error of the regression coefficient. Here, the regression coefficient we are working with is b1, so the standard error of this coefficient is denoted sb1.

It is too difficult to compute b1 by hand, but, fortunately, Minitab computes the standard errors of all the regression coefficients for us.

From the output we see that b_1 is 0.0811. Once we have found sb1 on the output, we can use Equation following to standardize the slope of our fitted regression equation:

Standardized Regression Coefficient

$$t = \frac{b_i - B_{i_0}}{S_{b_i}}$$

where

- b_i slope of fitted regression.
- B_{i_0} actual slope hypothesized for the population.
- S_{b_1} standard error of the regression coefficient.

In our example, the standardized value of the regression coefficient is

$$t = \frac{b_i - B_{i_0}}{S_{b_i}}$$

$$= \frac{0.597 - 0.400}{0.081}$$

$$= 2.432$$

Standardized Regression Coefficient

Conducting the hypothesis test

Suppose we are interested in testing our hypothesis at the 10 percent level of significance. Because we have 10 observations in our sample data, and three independent variables, we know that we have n – k –1 or 10 – 3 – 1 = 6 degrees of freedom.

The standardized regression coefficient is 2.432, which is outside the acceptance region for our hypothesis test. Therefore, we reject the null hypothesis that B1 still equals 0.400.

Confidence interval for Bi

In addition to hypothesis testing, we can also construct a confidence interval for any one of the values of Bi. In the same way that bi is a point estimate of Bi, such confidence intervals are interval estimates of Bi. To illustrate the process of constructing a confidence interval, let's find a 95 percent confidence interval for B_3 in our IRS problem. The relevant data are

$b_3 = 0.405$

$S_{b_3} = 0.0422$

t = 2.447 ← 5 percent level of significance and 6 degrees of freedom With this information, we can calculate confidence intervals like this:

$b_3 + t(S_{b_3}) = 0.405 + 2.447\ (0.0422)$

$= 0.508 \leftarrow$ Upper limit

$b_3 - t(S_{b_3}) = 0.405 - 2.447\ (0.0422)$

$= 0.302 \leftarrow$ Lower limit

Test of Whether a Variable Is Significant

$-tc \leq to \leq tc$

where

- tc = appropriate t value (with n "k "1 degrees of freedom) for the significance level of the test
- to = b /s i bi = observed (or computed) t value obtained from computer output

If to falls between "tc and tc, we accept H0 and conclude Xi is not a significant explanatory variable. Otherwise, we reject H0 and conclude that Xi is a significant explanatory variable.

Testing the significance of computer hours in the IRS problem.

Let's test, at the 0.01 significance level, whether computer hours is a significant explanatory variable for unpaid taxes discovered. with in -k -1 =10 -3 -1 =6 degrees of freedom and α =0.01, we see that tc = 3.707. Because t0 > tc, we conclude that computer hours is a significant explanatory variable. In fact, looking at the computed t values for the other two independent variables (field-audit labor hours t0 =7.36 and rewards to informants, t0 =9.59), we see that each of them is also a significant explanatory variable.

Q26. Explain Inferences about the Regression as Whole using an F test for population parameters.

Ans :

Inferences about the Regression as a Whole (Using an F Test).

Significance of the regression as a whole

Given any simple (or multiple) regression, it's natural to ask whether the value of r2 (or R2) really indicates that the independent variables explain Y, or might have happened just by chance. Our hypotheses are

H0: B1 B2 =... =Bk=0

Null hypothesis: Y doesn't depend on the Xi's.

H1: at least one Bi $\neq$ 0

Alternative hypothesis: Y depends on at least one of the Xi's.

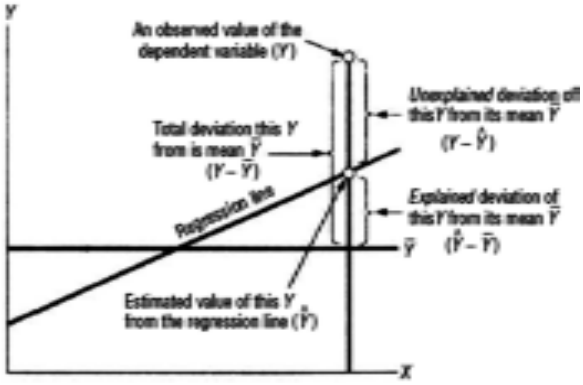


Fig.: Total Deviation, Explained Deviation, and Unexplained Deviation for one Observed Value of Y

Analyzing the variation in the Y values

We looked at the total variation in Y, Σ(Y -Y')2, the part of that variation that is explained by the regression $\Sigma(\hat{Y} - Y')2$, and the unexplained part of the variation, $\Sigma(Y - \hat{Y})2$. Above Figure reviews the relationship between total deviation, explained deviation, and unexplained deviation for

a single data point in a simple regression n. Although we can not draw a similar picture for multiple regression, we are doing the same thing conceptually.

Sums of squares and their degrees of freedom

Three Different Sums of Squares

SST = Total sum of squares (i.e., the explained part) = $\Sigma(Y - \bar{Y})^2$.

SSR = Regression sum of squares (i.e., the explained part) = $\Sigma(\hat{Y} - \bar{Y})^2$.

SSE = Error sum of squares (i.e., the explained part) = $\Sigma(\hat{Y} - \bar{Y})^2$.

Decomposing the Total Variation in Y

$$SST = SSR + SSE$$

F test on the regression as a whole

Each of these sums of squares has an associated number of degrees of freedom. SST has n – 1 degrees of freedom (n observations, less 1 degree of freedom because the sample mean is fixed). SSR has k degrees of freedom because there are k independent variables being used to explain Y. Finally, SSE has n – k – 1 degrees of freedom because we used our n observations to estimate k = 1 constants, a, b1, b2, ...,bk. If the null hypothesis is true, the ratio below has an F distribution with k numerator degrees of freedom and n -k -1 denominator degrees of freedom.

F Ratio

$$F = \frac{SSR / k}{SSE / (n - k - 1)}$$

If the null hypothesis is false, then the *F* ratio tends to be larger than it is when the null hypothesis is true. So if the *F* ratio is too high we reject H0 and conclude that the regression as a whole *is* significant.

Analysis of variance for the regression

This part of the output includes the computed F ratio for the regression, and is sometimes called the analysis of variance (ANOVA) for the regression..

For the IRS problem, we see that SSR = 29.109 (with k – 3 degrees of freedom), SSE = 0.491 (withn – k –1 = 10 – 3 – 1 = 6 degrees of freedom), and that,

$$F = \frac{29.109 / 3}{0.491 / 6} = \frac{9.703}{0.082} = 118.33$$

FACULTY OF INFORMATICS

M.C.A. I Year I Semester Examination

Model Paper - I

PROBABILITY AND STATISTICS

Time : 3 Hours] Max. Marks : 70

Answer all the question according to the internal choice **(5 × 14 = 70)**

ANSWERS

1. (a) Let H be the set of all vector of the form $\begin{bmatrix} 2t \\ 0 \\ -t \end{bmatrix}$ show that. H is subspace of R^3. **(Unit-I, Prob. 1)**

   (b) $A = \begin{bmatrix} 1 & -3 & -2 \\ -5 & 9 & 1 \end{bmatrix}$ $U = \begin{bmatrix} 5 \\ 3 \\ -2 \end{bmatrix}$ Determine if U belongs to the Null space of A. **(Unit-I, Prob. 7)**

2. (a) Define Kernel and Range of a linear transformation. **(Unit-I, Q.No. 6)**

   (b) Find bases for the null spaces of the matrix. $\begin{bmatrix} 1 & 0 & -3 & 2 \\ 0 & 1 & -5 & 4 \\ 3 & -2 & 1 & -2 \end{bmatrix}$. **(Unit-I, Prob. 21)**

3. (a) Explain the basic terminology are used in probability. **(Unit-II, Q.No. 2)**

   (b) In a sample of 446 cards, stopped at a road block, only 67 of the drivers, has their seat belts fastered. Estimate the probability that a driver stopped on that road, will have his or her seat belt fastered. **(Unit-II, Prob. 3)**

4. (a) In a certain college, 25% of boys and 10% of girls are studying mathematics. The girls constitute 60% of the student body.

   (i) What is the probability that mathematics is being studied ?

   (ii) If a student is selected at random and is found to be studying mathematics, find the probability that the student is a girl?

   (iii) a boy ? **(Unit-II, Prob. 9)**

   (b) Define Random Variables. Explain different tyeps of Random Variables. **(Unit-II, Q.No. 9)**

5. (a) What are the properties of a good estimator? **(Unit-III, Q.No. 11)**

   (b) In a study of an automobile insurance a random sample of 80 body repair costs had a mean of Rs. 472.36 and the S.D of Rs. 62.35. If $\bar{x}$ is used as a point estimate to the true average repair costs, with confidence we can assert that the maximum error doesn't exceed Rs. 10. **(Unit-III, Prob. 1)**

6. (a) Explain briefly about various types of estimates. **(Unit-III, Q.No. 10)**

(b) A random sample of size 16 as 53 mean the sum of the squares of the deviation taken from the mean is 135. Can these sample regarded as taken from the population having 56 mean ? Obtain at 1% Los $t_{0.01} = 2.95$. **(Unit-III, Prob. 6)**

7. (a) Define Hypothesis. Explain the procedure for testing a Hypothesis. **(Unit-IV, Q.No. 1)**

(b) Blood glucose levels for obese patients have a mean of 100 with a standard deviation of 15. A researcher thinks that a diet high in raw cornstarch will have a positive or negative effect on blood glucose levels. A sample of 30 patients who have tried the raw cornstarch diet have a mean glucose level of 140. Test the hypothesis that the raw cornstarch had an effect. **(Unit-IV, Prob. 2)**

8. (a) Write about hypothesis testing - single proportion. **(Unit-IV, Q.No. 6)**

(b) In a city A, out of 600 men, 325 men were found to be smokers. Does this information support the statement that 'Majority of men is this city are smokers'? Draw your conclusion using a = 1% and a = 5% separately. **(Unit-IV, Prob. 5)**

9. (a) Define Chi-square test. **(Unit-V, Q.No. 1)**

(b) Four different drugs have been developed for a certain disease. These drugs are used under three different environments. It is assumed that the environment might affect efficacy of drugs. The number of cases of recovery from the disease per 100 people who have taken the drugs is tabulated as follows :

Environment	Drug A1	Drug A2	Drug A3	Drug A4
I	19	8	23	8
II	10	9	12	6
III	11	10	13	16

Test whether the drugs differ in their efficacy to treat the disease, also whether there is any effect of environment on the efficacy of disease. **(Unit-V, Prob. 7)**

10. (a) Explain briefly about F-Distribution. **(Unit-V, Q.No. 9)**

(b) Determine the equation of a straightline which best fits the data.

X :	10	12	13	16	17	20	25
Y :	10	22	24	27	29	33	37

(Unit-V, Prob. 17)

FACULTY OF INFORMATICS

M.C.A. I Year I Semester Examination

Model Paper - II

PROBABILITY AND STATISTICS

Time : 3 Hours] Max. Marks : 70

Answer all the question according to the internal choice **(5 × 14 = 70)**

ANSWERS

1. (a) Let $A = \begin{bmatrix} 2 & 4 & -2 & 1 \\ -2 & -5 & 7 & 3 \\ 3 & 7 & -8 & 6 \end{bmatrix}$

(i) If the column space of A is a subspace of R^k, what is k?

(ii) If the Null space of A is a subspace of R^k what is k? **(Unit-I, Prob. 9)**

(b) Verify the set W = {(x, y, 0)/x, y $\in$ F)} forms a subspace of $V_3(F)$. **(Unit-I, Prob. 3)**

2. (a) Find A such that the set $\left\{\begin{bmatrix} b-c \\ 2b+3d \\ b+3c-3d \\ c+d \end{bmatrix}\right\}$ is Col A **(Unit-I, Prob. 15)**

(b) Suppose $\{\alpha_1\alpha_2,.....\alpha_K\}$ is L.D spanning set for a vector space V. Show that each W in V can be expressed in more then one way as a linear combination of $\alpha_1,\alpha_2.....\alpha_k$. **(Unit-I, Prob. 28)**

3. (a) Explain addition theorem of probability. **(Unit-II, Q.No. 4)**

(b) The probability that a contractor will get a plumbing contract is 3/4 and the probability that he will not get electric contract is 4/9. If the probability of getting at least one contract is 5/6, what is the probability that he will get both the contracts? **(Unit-II, Prob. 5)**

4. (a) A bag A contains 2 white and 3 red balls and a bag B contains 4 white and 5 red balls. One ball is drawn at random from one of the bags and it is found to be red. Find the probability that the red ball drawn is from bag B. **(Unit-II, Prob. 11)**

(b) A random variable X has the following probability function :

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

(i) Determine K

(ii) Evaluate P(X < 6), P(X $\geq$ 6), P(0 < X < 5) and P(0 $\leq$ X $\leq$ 4)

(iii) if P(X $\leq$ K) > $\frac{1}{2}$, find the minimum value of K and,

(iv) Determine the distribution function of X

(v) Mean

(vi) Variance **(Unit-II, Prob. 16)**

5. (a) Explain the term Sampling. **(Unit-III, Q.No. 1)**

(b) Assuming that $\sigma = 20.0$, how large a random sample be taken to assert with probability 0.95 that the sample mean will not differ from the true mean by more that 3.0 points ? **(Unit-III, Prob. 2)**

6. (a) Define the term Estimation. **(Unit-III, Q.No. 9)**

(b) Ten workers were given a training programme with a view to study their assembly time for a certain mechanism. The results of the time and motion studies before and after the training programme are given below

Workers	1	2	3	4	5	6	7	8	9	10
X_1	15	18	20	17	16	14	21	19	13	22
Y_1	14	16	21	10	15	18	17	16	14	20

X_1 = Time taken for assembling before training.

Y_1 = Time taken for assembling after training.

Test whether there is significant difference in assembly time before and after training. **(Unit-III, Prob. 8)**

7. (a) Explain the types of errors in sampling. **(Unit-IV, Q.No. 2)**

(b) In a simple random ample of 600 men taken from a big city 450 are found to be the users of a product 'A'. In another simple random sample of 900 men taken from another city 450 are found to be the users of a product 'A'. Do the data indicate that there is a significant difference in the habit of usage of the product 'A' in the two cities? **(Unit-IV, Prob. 6)**

8. (a) Explain briefly about Hypothesis Concerning two Proportion. **(Unit-IV, Q.No. 13)**

(b) A manufacturer of pet animal foods was wondering whether cat owners and dog owners reacted differently to premium pet foods. They commissioned a consumer survey that yielded the following data.

Pet	Number of owners Surveyed	Number of owners using Premium food
Cat	280	152
Dog	190	81

is it reasonable to conclude at a = 0.05, the cat owners are more likely to feed their pets. Premium food than dog owners? **(Unit-IV, Prob. 7)**

9. (a) "Chi-square act has a Test of Independence". Elaborate. **(Unit-V, Q.No. 4)**

(b) Find the coefficient of correlation between x and y

X :	65	66	67	67	68	69	70	72
Y :	67	68	65	68	72	72	69	71

(Unit-V, Prob. 12)

10. (a) Define Correlation. Explain different types of Correlation. **(Unit-V, Q.No. 10)**

(b) From the following data obtain the two regression equations :

X :	6	2	10	4	8
Y :	9	11	5	8	7

(Unit-V, Prob. 15)

FACULTY OF INFORMATICS

M.C.A. I Year I Semester Examination

Model Paper - III

PROBABILITY AND STATISTICS

Time : 3 Hours] Max. Marks : 70

Answer all the question according to the internal choice **(5 × 14 = 70)**

ANSWERS

1. (a) Find a matrix A such that

$W = \text{Col}A$ where

$$W = \left\{ \begin{bmatrix} 6a - b \\ a + b \\ -7a \end{bmatrix} \right\} \text{ a, b in R.}$$ **(Unit-I, Prob. 8)**

(b) $H = \text{Span}\{V_1, V_2, V_3\}$ & $B = (V_1\ V_2\ V_3)$. Show that B is a basis for H & X is in H, find the B-coordinate vector of X fee.

$$V_1 = \begin{bmatrix} -6 \\ 4 \\ -9 \\ 4 \end{bmatrix} V_2 = \begin{bmatrix} 8 \\ -3 \\ 7 \\ -3 \end{bmatrix} V_3 = \begin{bmatrix} -9 \\ 5 \\ -8 \\ 3 \end{bmatrix} X = \begin{bmatrix} 4 \\ 7 \\ -8 \\ 3 \end{bmatrix}$$ **(Unit-I, Prob. 31)**

2. (a) $\alpha = (1, 0, -1)$ w.r.t basis $S = \{x, y, z\}$ where

$$x = \begin{bmatrix} 0 \\ 1 \\ -1 \end{bmatrix} \quad y = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} \quad z = \begin{bmatrix} 1 \\ 0 \\ 2 \end{bmatrix}$$ **(Unit-I, Prob. 34)**

(b) Explain about Vector Spaces and Subspaces. **(Unit-I, Q.No. 1)**

3. (a) A card is drawn from a standard pack of 52 cards.
   (i) What is the probability that it is a black or a red card?
   (ii) What is the probability that it is an Ace or a King or a Queen?
   (iii) What is the probability that it is a red Knave or a black King? **(Unit-II, Prob. 6)**

   (b) Write about the various probabilities under statistical independence. **(Unit-II, Q.No. 5)**

4. (a) State the explain Bayes' theorem. **(Unit-II, Q.No. 8)**

   (b) It is known from past experience that in a certain industrial plant there are on the average 4 industrial accidents per month. Find the probability that in a given year there will be less than 4 accidents. Assume Poisson distribution ($e^{-4} = 0.0183$) **(Unit-II, Prob. 20)**

5. (a) Explain the determination of Sample Size in Estimation. **(Unit-III, Q.No. 14)**

   (b) It is desired to estimate the mean number of hours of continuous use until a certain computer will first require repairs. If it can be assumed that σ = 48 hours, how large a sample be needed so that one will be able to assert with 90% confidence that the sample mean a off by at most 10 hours. **(Unit-III, Prob. 3)**

6. (a) Explain different types of sampling methods. **(Unit-III, Q.No. 5)**

(b) Score obtained in shooting competition by 10 soldiers before and after intensive training are given below :

Before	67	24	57	55	63	54	56	68	33	43
After	70	38	58	58	56	57	68	75	42	38

Test whether the intensive training is useful at 0.05 level of significance. **(Unit-III, Prob. 9)**

7. (a) Elaborate the various Powers of hypothesis test. **(Unit-IV, Q.No. 5)**

(b) The mean lifetime of a sample of 100 light tubes produced by a company is found to be 1,560 hours with a population standard deviation of 90 hours. Test the hypothesis that the mean lifetime of the tubes produced by the company is 1580 hours. **(Unit-IV, Prob. 8)**

8. (a) Write about Hypothesis testing of two means. **(Unit-IV, Q.No. 9)**

(b) The Financial Accounting Board (FAB) was considering a proposal to require companies to report the potential effect of employees' stock options on earnings per share (EPS). A random sample of 41 high-technology firms revealed that the new proposal would reduce EPS by an average of 13.8 per cent, with a standard deviation of 18.9 per cent.

A random sample of 35 producers of consumer goods showed that the proposal would reduce EPS by 9.1 per cent on average with a standard deviation of 8.7 percent. On the basis of these samples, can it be reasonable to conclude at a = 0.05 that the FAB proposal will causes a greater reduction in EPS for high-technology firms than for producers of consumer goods? **(Unit-IV, Prob. 12)**

9. (a) What is ANOVA? What are its assumptions and applications? **(Unit-V, Q.No. 6)**

(b) Test whether the significance of possible variation in performance in a certain test between the grammar schools of a City A common test is given to number of students taken of random from the senior Vth Class of four schools.

Sl. No.	Schools			
	A	B	C	D
1	8	12	18	13
2	10	11	12	9
3	12	9	16	12
4	8	14	6	16
5	7	4	8	15

(Unit-V, Prob. 4)

10. (a) What is Karl Pearson's Coefficient of Correlation? Explain properties of Coefficient of Correlation. **(Unit-V, Q.No. 13)**

(b) How to find multiple regression equations. **(Unit-V, Q.No. 24)**

FACULTY OF INFORMATICS

M.C.A (CBCS) I - Semester Examination

August - 2021

PROBABILITY AND STATISTICS

Time : 2 Hours] **[Max. Marks : 70**

PART - A - ($4 \times 17^{1/2} = 70$ Marks)

Note : Answer any four Questions.

1. (a) Define sub space and Column space. **(Unit-I, Q.No. 1, 4)**

   (b) H = Span $\{V_1, V_2, V_3\}$ & B = (V_1, V_2, V_3) Show that B is a basis for H & X is in H, find the B-coordinate vector of X fee.

$$V_1 = \begin{bmatrix} -6 \\ 4 \\ -9 \\ 4 \end{bmatrix} \quad V_2 = \begin{bmatrix} 8 \\ -3 \\ 7 \\ -3 \end{bmatrix} \quad V_3 = \begin{bmatrix} -9 \\ 5 \\ -8 \\ 3 \end{bmatrix} \quad X = \begin{bmatrix} 4 \\ 7 \\ -8 \\ 3 \end{bmatrix}$$ **(Unit-I, Q.No. 31)**

2. (a) Define Linear Transformation and Basis. **(Unit-I, Q.No. 5)**

   (b) Let $B = \left\{ \begin{bmatrix} 1 \\ -4 \end{bmatrix} \begin{bmatrix} -2 \\ 9 \end{bmatrix} \right\}$ Since the coordinate mapping determined by B is a linear trans formation from R^2 into R^2 this mapping must be implanted by some 2×2 matrix A find A. **(Unit-I, Q.No. 32)**

3. In a single throw with two dice find the probability of throwing a sum

   (i) 10

Ans :

Sum Of Observation = 36 [(1,1),(1,2),(1,3),(1,4),(1,5),(1,6)

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

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

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

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

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

Favourable Observation = 3[(4,6),(5,5),(6,4)]

Probability = Favourable Observations / No of Observations

P(of getting Numbers Whose Sum = 10) = 3/36

P(of getting Numbers Whose Sum = 10) = 1/12

(ii) Which is a perfect square.

Ans :

Total no. of outcomes = 36

Let E be the event of getting the sum as perfect square.

No. of outcomes favourable to E = 7 {(1,3) (2,2) (3,1) (3,6) (4,5) (5,4) (6,3)}

Probability of E = No. of outcomes favourable to E / Total No. of Outcomes

= 7 / 36

4. If X is a normal variate with mean and standard deviation 5. Find the probabilities that (i) $26 \leq x \leq 40$ and (ii) $x \geq 45$

Ans :

According to the question, X is the standard normal variate with mean $\mu = 5$ and standard deviation $\sigma = 5$.

So :

X ~ N(5, 25)

(i) Now, the probability $P(26 \leq X \leq 40)$ can be calculated in the following manner :

$$P(26 \leq X \leq 40) = P\left[\frac{26-\mu}{\sigma} \leq \frac{X-\mu}{\sigma} \leq \frac{40-\mu}{\sigma}\right]$$

$$= P\left[\frac{26-5}{5} \leq Z \leq \frac{40-5}{5}\right]$$

$$= P(4.2 \leq Z \leq 7)$$

$$= P(Z \leq 7) - P(Z \leq 4.2)$$

$$= 1 - 0.999986654$$

$$= 0.000013346$$

Here, $Z = \frac{X-\mu}{\sigma}$ is the standard normal variate.

(ii) Now, the probability $P(X \geq 45)$ can be calculated in the following manner.

$$P(X \geq 45) = P\left[\frac{X-\mu}{\sigma} \geq \frac{45-\mu}{\sigma}\right]$$

$$= P\left[Z \geq \frac{45-5}{5}\right]$$

$$= P(Z \geq 8)$$

$$= 1 - P(Z < 8)$$

$$= 1 - 1$$

$$= 0$$

5. A population consists of six numbers 4, 8, 12, 16, 20, 24. Consider all samples of size two which can be drawn without replacement from this population. Find (i) the population mean and (ii) the population standard deviation

Ans :

The population mean is the average of all possible samples of a given size. For a sample of size two without replacement, you'll consider all possible pairs of numbers.

$\mu = \Sigma$ All possible pairs /Number of pairs

In this case:

$\mu = (4+8)+(4+12)+(4+16)+(4+20)+(4+24)+(8+12)+(8+16)+(8+20)+(8+24)+(12+16)+(12+20)+(12+24)+(16+20)+(16+24)+(20+24)/15$

$=28$

Population Standard Deviation (σ):

The population standard deviation is a measure of the amount of variation or dispersion in a set of values. For a sample of size two without replacement, you'll calculate the standard deviation using the formula:

The population standard deviation is a measure of the amount of variation or dispersion in a set of values.

$\sigma = \Sigma(X_i - \mu)^2 / N$

$=(4-28)^2+(8-28^2)+(12-28)^2+(16-28)^2+(20-28)^2+(24-28)^2/5$

$=17.0645832$

6. (a) The mean height of students in a college is 155 cms and standard deviation is 1.5. What is the probability that the mean height of 36 students is less than 157 cms?

Ans :

Given,

Mean = μ = 155

S.d = σ = 15

n = 36

Let x' be the mean random variable that follows normal distribution then the formula to calculate the z score is

$$z = \frac{x' - \mu(x')}{\sigma(x')}$$

Where,

$\mu(x') = \mu = 155$

$$\sigma(x') = \frac{\sigma}{\sqrt{n}} = \frac{15}{\sqrt{36}} = 2.5$$

Then,

$$P(x' < 157) = P\left(\frac{x' - \mu(x')}{\sigma(x')} < \frac{157 - 155}{2.5}\right)$$

$$= P(Z < 0.8)$$

$$= 0.78814$$

(b) A random sample of size 100 has a standard deviation of 5. What can you say about the maximum error with 95% confidence?

Ans :

The critical value for $\alpha = 0.05$ is $z_c = z_{1-\alpha/2} = 1.96$.

Maximum error $E = z_c \times \frac{\sigma}{\sqrt{n}}$

$$E = 1.96 \times \frac{5}{\sqrt{100}} = 0.98$$

---

7. A sample of 400 items is taken from a population whose standard deviation is 10. The mean of the sample is 40. Test whether the sample has come from a population with mean 38. Also calculate 96% confidence in interval for the population.

Ans :

Given

- Population standard deviation (σ) = 10
- Population mean under the null hypothesis (μ) = 38
- Sample mean $\bar{x} = 40$
- Sample size (n) = 400

Objectives

- 96% Confidence interval
- Test whether the sample comes from a population with a mean of 38.

Confidence Interval

The formula for the 96% Confidence interval is :

Confidence Interval = $\bar{X} \pm Z_{\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}}$

Where :

- $\bar{X}$ is the sample mean,

- $Z_{\frac{\alpha}{2}}$ is the critical value, For a 96% confidence interval, the critical value of $Z_{\frac{0.04}{2}} = Z_{0.02}$ is 2.05
- σ is the population standard deviation,

Setting up of the hypotheses:

$H_0 : \mu = 38$

$H_1 : \mu \neq 38$ (Two Tailed Test)

The decision rule for a hypothesis test based on a confidence interval is as follows:

Hypothesis Test Decision Rule :

- If the null hypothesis (H_0) falls outside the confidence interval :
- Reject the null hypothesis
- If the null hypothesis (H_0) falls inside the confidence interval:
- Fail to reject the null hypothesis.

Calculation of 96% Confidence interval for Mean :

$$CI = 40 \pm 2.05 \times \left(\frac{10}{\sqrt{400}} \right)$$

$$= 40 \pm 2.05 \times \frac{10}{20}$$

$$= 40 \pm 2.05 \times 0.5$$

$$= 40 \pm 1.025$$

The lower and upper limits of the confidence interval are :

Lower Limit = 40 – 1.025

= 38.975

Also,

Upper Limit = 40 + 1.025

= 41.025

Therefore, the confidence interval is : (38.975, 41.025)

Since 38 falls outside the interval (38.975, 41.025), we reject the null hypothesis at the 96% confidence level. This suggests that there is enough evidence to conclude that the sample did not come from a population with a mean of 38.

8. A researcher wants to know the intelligence of students in a school. He selected two groups of students. In the first group there are 150 student having mean 1Q of 75 with a S.D. of 15, in the second group there are 250 students having mean 1Q of 70 with S.D. of 20.

Ans :

Standard deviation $s_1 = 15$

Total number of students $n_2 = 250$

Mean $\bar{X}_2 = 70$

Standard deviation $s_2 = 20$

(a) The Null hypothesis H_0 :

The sample has been drawn from a sample population.

i.e : $\mu_1 = \mu_2$

Alter native hypothesis H_1 :

$\mu_1 \neq \mu_2$

Test statistic = $Z \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}}$

$= \frac{75 - 70}{\sqrt{\frac{(15)^2}{150} + \frac{(20)^2}{250}}}$

$= \frac{5}{\sqrt{\left(\frac{225}{150} + \frac{400}{250}\right.}}$

$= \frac{5}{\sqrt{1.5 + 1.6}}$

$= 2.83980917124$

$Z = 2.84$

$Z_{\frac{\alpha}{2}}$ for 0.01 the level of significance = 2.58

$= Z > Z_{\frac{\alpha}{2}}$

The null hypothesis is rejected, and it is concluded that the groups have not come from the same population.

9. A firm manufacturing rivets wants to limit variations in their length as much as possible. The lengths(in cms) of 10 rivets manufactured by a new process are

2.15	1.99	2.05	2.12	2.17
2.01	1.98	2.03	2.25	1.93

Examine whether the new process can be considered superior to the old if the old population has standard deviation 0.145 cm?

Ans :

We have

$n = 10, \ \bar{x} = \frac{\Sigma x_i}{n} = \frac{20.68}{10} = 2.068$

$$s^2 = \frac{\Sigma(x_i - \bar{x})^2}{n} = \frac{0.09096}{10} = 0.0091$$

and $\sigma_0 = 0.145$

1. Null Hypothesis $H_0 : \sigma^2 = \sigma_0^2$
2. Alternative Hypothesis $H_1 : \sigma^2 > \sigma_0^2$
3. Level of significance, $\alpha = 0.05$
4. Assuming that H_0 is true, the test statistic is

$$\chi^2 = \frac{ns^2}{\sigma_0^2} = \frac{0.09096}{(0.145)^2} = 4.326 \quad [\because \sigma_0 = 0.145 \text{ (given)}]$$

$\therefore$ Calculated $\chi^2 = 4.8$

d.f. = n – 1 = 10 – 1 = 9

Tabulated χ^2 < tabulated χ^2, we accept the null hypothesis H_0. i.e., The new process cannot be considered superior to the old process.

10. Find if there is any significant correlation between the heights and weights given Below and

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

Ans :

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

Coefficient of correlation $r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \Sigma y^2}} = \frac{216}{\sqrt{(102)(471)}} = 0.98$

FACULTY OF INFORMATICS

MCA I-Semester (CBCS) (Main & Backlog) Examinations

April / May - 2023

PROBABILITY AND STATISTICS

Time : 3 Hours] **[Max. Marks : 70**

Note : I. Answer one questions from each unit. All questions carry equal marks.

II. Missing data, if any, may be sutiably assumed.

UNIT - I

1. (a) Explain about Vector Spaces and Subspaces. **(Unit-I, Q.No.1)**

(b) Prove the set of solutions (x, y, z) of the equations $x + y + 2z = 0$ is a subspace of the space $R^3(R)$.

Sol :

To prove that the set of solutions (x,y, z) of the equation $x + y + 2z = 0$ is a subspace of the vector space R^3. we need to verify that it satisfies the three properties that define a subspace:

1. The zero vector is in the subspace.
2. The subspace is dosed under vector addition.
3. The subspace is closed under scalar multiplication.

Property 1: The Zero Vector

The zero vector in R^3 is (0, 0,0). To check if this vector satisfies the equation $x + y + 2z – 0$,

we can simply plug in its values:

$$0 + 0 + 2(0) = 0$$

So, the zero vector (0, 0, 0) is in the subspace.

Property 2: Closed under Vector Addition

To check if the subspace is closed under vector addition, we need to show that if two vectors (x_1, y_1, z_1) and (x_2, y_2, z_2) satisfy the equation $x + y + 2z = 0$, then their sum: should also satisfy the same equation:

Suppose $(x_1, y_1, z_1$ and (x_2, y_2, z_2) are in the subspace, which means they satisfy $x_1 + y_2 + 2z_1 = 0$ and $x_2 + y_2 + 2z_2 = 0$.

Now, consider their sum:

$$(x_1 + x_2) + (y_1 – y_2) + 2(z_1 + z_2)$$

Using the properties of real numbers and the equations for (x_1, y_1, z_1) and (x_2, y_2, z_2), we can write:

$$(x_1 + x_2 + (y_1 + y_2) + 2(z_1 + z_2) = (x_1 + y_1 + 2z_1) + (x_2 + y_2 + 2z_2) = 0 + 0 = 0$$

So, the sum of $(xi + x_2, yi + y_2, Zi + z_2)$ also satisfies the equation $x + y + 2z = 0$. Therefore, the subspace is dosed under vector addition.

Property 3: Closed under Scalar Multiplication

To check if the subspace is dosed under scalar multiplication, we need to show that if a vector (x, y, z) satisfies the equation x + y + 2z = 0, then any scalar multiple of that vector also satisfies the same equation.

Suppose (x, y, z) is in the subspace, which means it satisfies x + y + 2z = 0. Now, consider a scalar c :

c(x, y, z) = (cx, cy, cz)

Let's check if c(x, y, z) satisfies the equation:

cx + cy + 2cz

Using the properties of real numbers and the equation for (x. y, z), we can write:

cx + cy 4 2cz = c(x + y + 2z) = c(0) = 0

So, the scalar multiple c(x, y, z) also satisfies the equation x + y + 2z = 0. Therefore, the subspace is dosed under scalar multiplication.

OR

2. (a) Define column space of a (m × n) matrix A. **(Unit-I, Q.No.4)**

(b) Determine Whether set of polynomial form a basis of P_3.

$5 - 3t + 4t^2 + 2t^3$

$9 + t + 8t^2 - 6t^3$

$6 - 2t + 5t^2$

t^3

Sol :

Given polynomials

$5 - 3t + 4t^2 + 2t^3$

$g + t + 8t^2 - 6t^3$

$6 - 2t + 5t^2$

t^3

above polynomials can be rewritten as,

$5 + (-3) + 4t^2 + 2t^3$

$g + 1t + 8t^2 + (-6)t^3$

$6 + (-2)t + 5t^2 + 0t^3$

$0 + 0t + 0t^2 + 1t^3$

The corresponding determinant of their coefficients

$$\begin{vmatrix} 5 & -3 & 4 & 2 \\ 9 & 1 & 8 & -6 \\ 6 & -2 & 5 & 0 \\ 0 & 0 & 0 & 1 \end{vmatrix}$$

Expanding about 4th row

$$= -0 + 0 - 0 + 1\begin{vmatrix} 5 & -3 & 4 \\ 9 & 1 & 8 \\ 6 & -2 & 5 \end{vmatrix}$$

$$= \begin{vmatrix} 5 & -3 & 4 \\ 9 & 1 & 8 \\ 6 & -2 & 5 \end{vmatrix}$$

$$= 5\begin{vmatrix} 1 & 8 \\ -2 & 5 \end{vmatrix} + 3\begin{vmatrix} 9 & 8 \\ 6 & 5 \end{vmatrix} + 4\begin{vmatrix} 9 & 1 \\ 6 & -2 \end{vmatrix}$$

$= 5(5 + 16) + 3(45 - 48) + 4(-18 - 6)$

$= 5(21) + 3(-3) + 4(-24)$

$= 105 - 9 - 96$

$= 105 - 105$

$= 0$

UNIT - II

3. (a) Explain types of Events.

Ans :

The different types of events in probability are:

1. Sure event
2. Impossible event
3. Independent event
4. Dependent event
5. Mutually exclusive event
6. Complementary event
7. Compound event
8. Exhaustive event
9. Simple event

All these types of events are explained below with the help of examples.

1. Sure Event

It is an event that always occurs when an experiment is conducted. For example, getting a tail when a coin is tossed. The probability of a sure event is 1.

Example: The probability of an event that has all outcomes of the experiment, i.e., sample space, is 1.

2. Impossible Event

If the probability of occurrence of an event is zero, then it is an impossible event.

Example: The event of getting 7 when a die is thrown is impossible. This is because the outcomes of throwing a die include {1, 2, 3, 4, 5, 6}.

3. Independent Event

When the outcome of the first event does not influence the outcome of the second event, those events are known as independent events.

Example: The event of getting a tail after tossing a coin and the event of getting a head when tossing another coin.

4. Dependent Event

When the outcome of the first event influences the outcome of the second event, those events are called dependent events.

Example: If we draw two coloured marbles from a bag and the first marble is not replaced before we draw the second marble, then the outcome of the second draw will depend on the outcome of the first draw.

5. Mutually Exclusive Event

These events cannot happen at the same time. They cannot occur at the same time.

Example: The events of getting head and tail are mutually exclusive while tossing a coin.

6. Complementary Event

For any event A, another event, A', shows the remaining elements of the sample space S. A' = S – A.

Example: Suppose the set of the first 10 natural numbers is a sample space, S = {1, 2, 3, 4, 5, 6, 7, 8, 9, 10} and A be the event of choosing an even number less than 10. So, A = {2, 4, 6, 8}

Thus, A' = S – A = {1, 3, 5, 7, 9, 10}

7. Compound Event

If an event has more than one sample point, it is termed as a compound event.

Example: If S = {1, 2, 3, 4, 5, 6} such that $E_1 = \{1, 3, 6\}$, $E_2 = \{2, 6\}$. Thus, E_1 and E_2 represents compound events.

8. Exhaustive Event

The events $E_1, E_2, \ldots\ldots, E_n$ are exclusive if $E_1 \cup E_2 \cup \ldots\ldots \cup E_n = S$, where S is the sample space.

Example: Suppose E_1 be the event of getting an even number and E_2 be the event of getting an odd number when throwing a die.

Here, $E_1 = \{2, 4, 6\}$, $E_2 = \{1, 3, 5\}$

$E_1 \cup E_2 = \{1, 2, 3, 4, 5, 6\} = S$ (sample space)

Thus, E_1 and E_2 are exhaustive events.

9. Simple event

An event that has a single point of the sample space is known as a simple event in probability.

Example: If S = {1, 2, 3, 4} and E = {3} then E is a simple event.

(b) State and prove Baye's theorem. **(Unit-II, Q.No.8)**

OR

4. (a) What is Random Variable. **(Unit-II, Q.No.9)**

(b) Three cards are drawn at random successively with replacements from a well shuffled pack of cards. Getting a card of diamond is termed as a success. Obtain the probability distribution of the number of success.

Sol :

No. of diamonds = 13

No. of cards = 52

Probability of diamond = $\frac{1}{4}$

For 3 cards = $\left(\frac{1}{4}\right)^3 = \frac{1}{64}$

UNIT - III

5. (a) What is Random Sampling? **(Unit-III, Q.No.5)**

(b) The heights of 10 males of a given locality are found to be 70,67,62,68,61,68,70,64,64 and 66 inches. Is it reasonable to believe that the average height is greater than 64 inches.(Table value-1.833)

Sol :

Data of 10 males given in the question is,

70, 67, 62, 68, 61, 68, 70, 64, 64, 66

The objective of the question is to test the claim that the average height is greater than 64 inches.

Also $t_{0.05,9} = 1.833$ is given

Explanation

Null hypothesis is first requirement for doing this question. Null hypothesis is the hypothesis of no change.

Before testing Null and Alternative hypothesis we are required to find the sample mean and sample standard deviation for that consider the table given below.

x	c = X–mean of X	d ^ 2
70	4	16
67	1	1
62	-4	16
68	2	4
61	-5	25
68	2	2
70	4	16
64	-2	4
64	-2	4
65	0	0
Sum = 660	Sum = 0	Sum = 90

The average height of sample $\bar{x} = \frac{\sum x}{n} = \frac{660}{10} = 66$

The standard deviation of sample $s = \sqrt{\frac{\sum d^2}{n-1}} = \sqrt{\frac{90}{0}} = 3.16$

Null and Alternative hypothesis are,

$H_0 = \mu - 64$

$H_0 = \mu > 64$

Test statistics t is,

$$t = \frac{x - p}{\frac{n}{\sqrt{n}}}$$

Putting all values from the previous steps we get

$$t = \left(\frac{60 - 64}{\frac{3.16}{\sqrt{10}}} \right)$$

$$- \frac{2 \times \sqrt{10}}{3.16}$$

$= 2$

So, the calculated value of t is 2 and tabulated value given in the question is 1.833.

OR

6. (a) Define Interval Estimates. **(Unit-III, Q.No.10)**

(b) Explain Random and Non Random sampling. **(Unit-III, Q.No.5)**

UNIT - IV

7. (a) What is standard deviation?

Ans :

Standard Deviation is a measure which shows how much variation (such as spread, dispersion, spread,) from the mean exists. The standard deviation indicates a "typical" deviation from the mean. It is a popular measure of variability because it returns to the original units of measure of the data set. Like the variance, if the data points are close to the mean, there is a small variation whereas the data points are highly spread out from the mean, then it has a high variance. Standard deviation calculates the extent to which the values differ from the average. Standard Deviation, the most widely used measure of dispersion, is based on all values. Therefore a change in even one value affects the value of standard deviation. It is independent of origin but not of scale. It is also useful in certain advanced statistical problems.

(b) A Random sample of 40 student selected from a college and their average height is found to be 175 cms with a standard deviation of 10 cms. Another sample of 50 students selected from a college, their average height is found to be 170 cms with a standard deviation of 4 cms. Test an average, the average height of the students of 1st college is > 2nd college.

S0l :

Hypothesis

Null Hypothesis H_0 : The average height of students from the first college (p,) is less than or equal to the average height of students from the second college (P2).

H_0 : $\mu_1 \leq \mu_2$

Alternative Hypothesis H_1: The average height of students from the first college (pi) is greater than the average height of students from the second college (p_2).

H_1 : $\mu_1 > \mu_2$

Choose a significance level, which is the probability of making a Type I error (rejecting the null hypothesis when it is true). Common choices include a = 0.05 for a 95% confidence level.

For the first college:

Sample size $n_1 = 40$

Sample mean $\bar{x}_1 = 175$ cm

Sample standard deviation $s_1 = 10$ cm

For the second college:

Sample size $n_2 = 50$

Sample mean $\bar{x}_2 = 170$ cm

Sample standard deviation $S_2 = 4$ cm

Calculate the test statistic, which is the t-statistic, using the formula for a two-sample t- test:

$$t = \frac{(x_1 - x_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{n_2^2}{n_2}}}$$

$$t = \frac{175 - 170}{\sqrt{\left(\frac{10^2}{40}\right) + \left(\frac{4^2}{50}\right)}}$$

= 2.98

Critical value:

df = min(n_1 – 1, n_2 – l) if n_1 – 40 and n_2 = 50, then

cff = min(40 – l, 50 – l) = min(39,49) = 39

The critical value at a = 0.05 with df = 39 is 1.6849

t value < critical value. Hence we fail to reject null hypothesis and concluded that the average height of students from the first college (μ_1) is less than or equal to the average height of students from the second college (μ_2).

The average height of students from the first college (μ_1) is less than or equal to the average height of students from the second college (μ_2).

OR

8. (a) Explain large sample test.

Ans :

As a thumb rule, a sample of size n is treated as a large sample only if it contains more than 30 units (or observations, n > 30). And we know that, for large sample (n > 30), one statistical fact is that almost all sampling distributions of the statistic(s) are closely approximated by the normal distribution.

(b) Before increasing the exercise duty 40 out of 70 were smokers, after increasing the tax a random sample of 150 persons selected. In this sample 90 persons were found smokers. Test, is there any decrease in the proportion of smokers after increasing the tax.

Sol :

To test whether there is any decrease in the proportion of smokers after increasing the tax, you can perform a hypothesis test to compare proportions. Specifically, you can use a two-proportion z- test.

Let's set up the hypotheses for this test:

- **Null Hypothesis (H_0):** The proportion of smokers before increasing the tax (p_1) is the same as the proportion of smokers after increasing the tax (p_2), i. e., ($p_1 = p_2$).
- **Alternative Hypothesis (H_1):** The proportion of smokers before increasing the tax (p_1) is greater than the proportion of smokers after increasing the tax (p_2), i. e., ($p_1 > p_2$).

Now, let's calculate the test statistic using the given data:

- For the sample before the tax increase: ($n_1 = 70$), ($x_1 = 40$)
- For the sample after the tax increase: ($n_2 = 150$), ($x_2 = 90$)

The test statistic for the two-proportion z-test can be calculated as follows:

$$\left[z = \frac{\left(\hat{P}_1 - \hat{P}_2\right)}{SE} \right]$$

Here :

- $(\hat{P}_1)$ and $(\hat{P}_2)$ are the sample proportions for the two samples.
- $(\hat{P})$ is the overall sample proportion, calculated as $\left(\frac{x_1 + x_2}{n_1 + n_2}\right)$
- SE is the standard error (SE) of the difference in proportions which is calculated as :

$$SE = \sqrt{\hat{P}\left(1-\hat{P}\right)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

Let's calculated the test statistic :

$$\left(\hat{P}_1 = \frac{x_1}{n_1} = \frac{40}{70} = 0.571428571\right)$$

$$\left(\hat{P}_2 = \frac{x_2}{n_2} = \frac{90}{150} = 0.6\right)$$

$$\left(\hat{P} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{40 + 90}{70 + 150} = \frac{130}{220} = 0.590909091 \approx 0.5909\right)$$

Now, calculate the standard error (SE) of the difference in proportions :

$$SE = \sqrt{\hat{P}\left(1 - \hat{P}\right)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

$$SE = \sqrt{0.5909\left(1 - 0.5909\right)\left(\frac{1}{70} + \frac{1}{150}\right)}$$

$$SE = \sqrt{0.24173719\left(\frac{1}{70} + \frac{1}{150}\right)}$$

$$SE = \sqrt{0.2417\left(0.02095\right)}$$

$SE = \sqrt{0.005063615} = 0.071159082 \approx 0.07116$

Now, calculate the z-statistic :

$$\left[z = \frac{\left(\hat{P}_1 - \hat{P}_2\right)}{SE}\right]$$

$$z = \frac{0.571428571 - 0.6}{0.07116} = -\frac{0.028571429}{0.071159082}$$

$z = 0.401514862 \approx -0.4015$

The p-value associated with $z = -0.4015$ is $0.688052 \approx 0.6881$

With a p-value of 0.6881, you would typically compare it to your chosen significance level (alpha) to make a decision about the null hypothesis.

Since the p-value is much greater than alpha (0.6881 > 0.05 or any reasonable alpha level), you do not have sufficient evidence to reject the null hypothesis.

UNIT - V

9. (a) Define :

(i) Population (ii) Sample (iii) Parameter (iv) Statistics

(v) Standard error

Ans :

(i) Population

In statistics, a population is the pool of individuals from which a statistical sample is drawn for a study. Thus, any selection of individuals grouped by a common feature can be said to be a population. A sample may also refer to a statistically significant portion of a population, not an entire population.

(ii) Sample

A sample refers to a smaller, manageable version of a larger group. It is a subset containing the characteristics of a larger population. Samples are used in statistical testing when population sizes are too large for the test to include all possible members or observations. A sample should represent the population as a whole and not reflect any bias toward a specific attribute.

(iii) Parameter

A parameter is a number describing a whole population (e.g., population mean), while a statistic is a number describing a sample (e.g., sample mean).

(iv) Statistics

Statistics is a branch that deals with every aspect of the data. Statistical knowledge helps to choose the proper method of collecting the data and employ those samples in the correct analysis process in order to effectively produce the results. In short, statistics is a crucial process which helps to make the decision based on the data.

(v) Standard error

The standard error (SE) of a statistic is the approximate standard deviation of a statistical sample population.

The standard error is a statistical term that measures the accuracy with which a sample distribution represents a population by using standard deviation. In statistics, a sample mean deviates from the actual mean of a population; this deviation is the standard error of the mean.

(b) A survey of 320 families with 5 children each revealed the following distribution:

No. of boys:	5	4	3	2	1	0
No. of girls:	0	1	2	3	4	5
No. of Families:	14	56	110	68	40	12

Is this result consistent with the hypothesis that male and female births are equally probable.

Sol :

Step 1: Define the Hypotheses

- Null Hypothesis (H_0): Male and female births are equally probable.
- Alternative Hypothesis (H_1): Male and female births are not equally probable.

Step 2: Set the Significance Level

We set the significance level (a) to 0.05.

Step 3; Create Expected Frequencies

Explanation:

Assuming equal probability for boys and girls in each family:

Expected boys per family = 5 children $\times \left(\frac{1}{2}\right)$ = 2.5 boys

Expected girls per family = 5 children × $\left(\frac{1}{2}\right)$ = 2.5 girls

Step 4: Calculate the Chi-Square Statistic

Using the Chi-Square formula:

$$\chi = \Sigma\left[\frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}\right]$$

Calculating for each category:

$$\chi^2 = \left[\frac{(5-2.5)^2}{2.5}\right] + \left[\frac{(4-2.5)^2}{2.5}\right] + \left[\frac{(3-2.5)^2}{2.5}\right] + \left[\frac{(2-2.5)^2}{2.5}\right] + \left[\frac{(1-2.5)^2}{2.5}\right] + \left[\frac{(0-2.5)^2}{2.5}\right]$$

$$\chi^2 = \left[\frac{2.5^2}{2.5}\right] + \left[\frac{1.5^2}{2.5}\right] + \left[\frac{0.5^2}{2.5}\right] + \left[\frac{0.5^2}{2.5}\right] + \left[\frac{1.5^2}{2.5}\right] + \left[\frac{2.5^2}{2.5}\right]$$

$\chi^2 = 7.2$

Step 5: Determine Degrees of Freedom

Degrees of freedom (df) = Number of Categories –1 = 6 – 1 = 5.

Step 6: Find the Critical Value

Using a Chi-Square table or calculator with 5 degrees of freedom and α = 0.05, the critical value is approximately 11.0705.

Step 7: Compare the Chi-Square Statistic and Critical Value

The calculated Chi-Square statistic (χ^2 = 7.2) is less than the critical value (11.0705)

Step 8: Make a Decision

Since χ^2 (7.2) < critical value (11.0705), we fall to reject the null hypothesis.

Therefore, the result is consistent with the hypothesis that male and female births are equally probable.

OR

10. (a) The following are the regression lines 8x – 10y + 66 = 0 and 40x – 18y = 214.

Find:

i) Regression equation Y on X.

ii) Correlation coefficient between X and Y

Sol :

Y = a + bX

Where 'a' is the intercept and 'b' is the slope.

First, you need to express the given regression line equations in the form Y = a + bX :

1. $8x - 10y + 66 = 08x + 66 = l0y$

$$\frac{8x+66}{10} = y$$

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

So, the first regression equation is: $Y = \left(\frac{4}{5}\right)x + \frac{33}{5}$

2. $40x - 18y = 214$

$40x = 18y + 214$

$40x = 18y + 214$

$2x = 9y + 107$

$$\frac{2x-107}{9} = y$$

So, the second regression equation is: $Y = \left(\frac{2}{9}\right)x - \frac{107}{9}$

Correlation Coefficient between X and Y: The correlation coefficient (r) can be found using the slopes of the regression lines. The formula for r is:

$r = \pm\sqrt{b_1 \times b_2}$

Where b_1 and b_2 are the slopes of the two regression lines.

From the regression equations we found in part i:

For the first regression equation: $b_1 = \frac{4}{5}$ For the second regression equation: $b_2 = \frac{2}{9}$

Now, calculate the correlation coefficient:

$$r = \pm\sqrt{\left(\frac{4}{5}\right) \times \left(\frac{2}{9}\right)}$$

$$r = \pm\sqrt{\left(\frac{8}{45}\right)}$$

$$r = \pm\sqrt{\frac{8}{45}}$$

$r \approx \pm 0.7526$.

So, the correlation coefficient between X and Y is approximately ± 0.7526. The sign of the correlation coefficient indicates the direction of the relationship between X and Y (positive or negative), and the magnitude (absolute value) indicates the strength of the relationship.

(b) The following are the regression lines $8x - 10y + 66 = 0$ and $40x - 18y = 214$. Find:

i) Regression equation X on Y.

ii) Mean Values of X and Y

Regression line 1 : $8x - 10y + 66 = 0$

Rearrange it to $y = mx + c$ form:

$8x - 10y + 66 = 0$

$-10y = -8x - 66$

$$y = \left(\frac{8}{10}\right)x - \frac{66}{10}$$

$$= \left(\frac{4}{5}\right)x - \frac{33}{5}$$

So, the equation of the regression line X on Y is :

$$Y = \left(\frac{4}{5}\right)X - \frac{33}{5}$$

So, the equation of the regression line X on Y is :

$$Y = \left(\frac{4}{5}\right)X - \frac{33}{5}$$

Regression line 2 : $40x - 18y = 214$

Rearrange it to $y = mx + c$ form :

$40x - 18y = 214$

$-18y = -40x + 214$

$$y = \left(\frac{40}{18}\right)x - \frac{214}{18}$$

$$= \left(\frac{20}{9}\right)x - \frac{107}{9}$$

The mean value of X ($\bar{X}$) is calculated by summing all X values and dividing by the number of data points:

Mean of X ($\bar{X}$) = (Sum of X values) / (Number of data points)

The mean value of Y ($\bar{Y}$) is calculated in the same way:

Mean of Y ($\bar{Y}$) = (Sum of Y values) / (Number of data points)

However, in the context of regression equations, we don't have the actual data points to calculate the means because we only have the regression lines. To find the means of X and Y, you would need the actual data from which these regression lines were derived. If you have that data, you can calculate the means using the formulas mentioned above. Without the data, you won't be able to determine the means.

FACULTY OF INFORMATICS

MCA I - Semester (CBCS) (Backlog) Examinations

October / November - 2023

PROBABILITY AND STATISTICS

Time : 3 Hours] **[Max. Marks : 70**

Note : I. Answer one questions from each unit. All questions carry equal marks.

II. Missing data, if any, may be suitably assumed.

ANSWERS

Unit - I

1. a) Find a spanning set for the null space of the matrix $A = \begin{bmatrix} -3 & 6 & -1 & 1 & -7 \\ 1 & -2 & 2 & 5 & -1 \\ 2 & -4 & 5 & 8 & -4 \end{bmatrix}$.

Ans:

The given matrix is

$$M = \begin{bmatrix} -3 & 6 & -1 & 1 & -7 \\ 1 & -2 & 2 & 5 & -1 \\ 2 & -4 & 5 & 8 & -4 \end{bmatrix}$$

Have to spanning set for the null space of M.

We will first reduce the matrix and then will try to find the spanning set of the null space of M.

Operating

$R_2 \to 3R_2$

And, $R_3 \to 3R_3$

We have

$$M = \begin{bmatrix} -3 & 6 & -1 & 1 & -7 \\ 3 & -6 & 6 & 9 & -3 \\ 6 & -12 & 15 & 24 & -12 \end{bmatrix}$$

Operating

$R_2 \to R_2 + R_1$

And, $R_3 \to R_3 + 2R_1$

We have

$$\sim \begin{bmatrix} -3 & 6 & -1 & 1 & -7 \\ 0 & 0 & 5 & 10 & -10 \\ 0 & 0 & 13 & 26 & -26 \end{bmatrix}$$

Now operating

$R_3 \to 13R_2$

$R_3 \to 5R_3$

We have

$$\sim \begin{bmatrix} -3 & 6 & -1 & 1 & -7 \\ 0 & 0 & 65 & 130 & -130 \\ 0 & 0 & 65 & 130 & -130 \end{bmatrix}$$

Operating

$R_3 \to R_3 \to R_2$

and then operation

$$R_2 \to \left(\frac{1}{13}\right) R_2$$

we have

$$\sim \begin{bmatrix} -3 & 6 & -1 & 1 & -7 \\ 0 & 0 & 5 & 10 & -10 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Clearly Rank of the row reduced matrix is

rank (M) = 2

Therefore number of free variable of the system MX = 0 is "Number of columns - rank of M"

Here number of free variable = 5 – 2 = 3.

Let

$$X = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix}$$

Such that MX = O. That is

$$\begin{bmatrix} -3 & 6 & -1 & 1 & -7 \\ 0 & 0 & 5 & 10 & -10 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

Gives

$$-3x_1 + 6x_2 - x_3 + x_4 - 7x_5 = 0$$

$$5x_3 + 10x_4 - 10x_4 = 0$$

Since number of free variables is 3 let

$x_2 = r$

$x_4 = s$

$x_5 = t$

Then

$x_3 = -2s + 2t$

$x_1 = 2r + 3s - 9t$

Hence

$$X = r\begin{bmatrix}2\\1\\0\\0\\0\end{bmatrix} + s\begin{bmatrix}3\\0\\-2\\1\\0\end{bmatrix} + t\begin{bmatrix}-9\\0\\2\\0\\1\end{bmatrix}$$

Hence the required spanning set of the null space of M is

$$\left\{\begin{bmatrix}2\\1\\0\\0\\0\end{bmatrix}, \begin{bmatrix}3\\0\\-2\\1\\0\end{bmatrix}, \begin{bmatrix}-9\\0\\2\\0\\0\end{bmatrix}\right\}$$

---

b) Let H be the set of vectors of the form $\begin{bmatrix}3t\\0\\-7t\end{bmatrix}$, where t is real number. Show that H is a subspace of V.

Ans:

To demonstrate that (H) is a subspace of vector space (V), three requirements must be met:

1. Closure under vector addition

For any vectors v_1, v_2 in H, their sum $(v_1 + v_2)$ must also be in (H).

2. Closure under scalar multiplication

For any vector V in H and any scalar C, the product c. v must also be in H.

3. Contains the zero vector

The set H must contain the zero vector.

i) Closure under vector addition

Let $v_1 = \begin{bmatrix} 3t_1 \\ 0 \\ -7t_1 \end{bmatrix}$ and $v_2 = \begin{bmatrix} 3t_2 \\ 0 \\ -7t_2 \end{bmatrix}$ where t_1, t_2 are real numbers.

The sum $v_1 + v_2$ is :

$$v_1 + v_2 = \begin{bmatrix} 3t_1 + 3t_2 \\ 0 \\ -7t_1 - 7t_2 \end{bmatrix}$$

This is of the same form as H(i.e., $v_1 + v_2$) is of the form

$$v_1 = \begin{bmatrix} 3t \\ 0 \\ -7t \end{bmatrix}$$

ii) Closure under scalar multiplication

Let $v = \begin{bmatrix} 3t \\ 0 \\ -7t \end{bmatrix}$ and let C be a scalar.

The scalar product c.v is :

$$c \times v = c \times \begin{bmatrix} 3t \\ 0 \\ -7t \end{bmatrix} = \begin{bmatrix} 3t \\ 0 \\ -7t \end{bmatrix}$$

This is also of the same form as H so it is also a closed scalar multiplication.

iii) Contains the zero vector

The zero vector is $\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$ which is of the form $\begin{bmatrix} 3.0 \\ 0 \\ -7.0 \end{bmatrix}$

The set H contains the zero vector.

H satisfies all the three conditions it is a subspace of V.

(OR)

2. a) Determine if $[v_1, v_2, v_3]$ is Linearly dependent of Linearly independent, where

$$v_1 = \begin{bmatrix} 0 \\ 1 \\ 5 \end{bmatrix}, v_2 = \begin{bmatrix} 1 \\ 2 \\ 8 \end{bmatrix}, v_3 = \begin{bmatrix} 2 \\ 1 \\ 0 \end{bmatrix}$$

Ans:

We have given

$$v_1 = \begin{bmatrix} 0 \\ 1 \\ 5 \end{bmatrix}, v_2 = \begin{bmatrix} 1 \\ 2 \\ 8 \end{bmatrix}, v_3 = \begin{bmatrix} 2 \\ 1 \\ 0 \end{bmatrix}$$

We put $v_1 = a_1$, $v_2 = a_2$, $v_3 = a_3$

We know that,

$$\sum_{i=1}^{3} c_i a_i = 0$$

$$c_1 a_1 + c_2 a_2 + c_3 a_3 = 0$$

$$c_1 \begin{bmatrix} 0 \\ 1 \\ 5 \end{bmatrix} + c_2 \begin{bmatrix} 1 \\ 2 \\ 8 \end{bmatrix} + c_3 \begin{bmatrix} 2 \\ 1 \\ 0 \end{bmatrix} = 0$$

$c_2 + 2c_3 = 0$ ... (1)

$c_1 + 2c_2 + c_3 = 0$ ... (2)

$5c_1 + 8c_2 = 0$ ... (3)

Equation (2) – 2 × equation (1)

$$\begin{array}{r} c_1 + 2c_2 + c_3 = 0 \\ 2c_2 + 4c_3 \\ - \quad - \\ \hline c_1 - 3c_3 = 0 \end{array} \Rightarrow c_1 = 3c_3$$

4 × equation (2) – equation (3)

$$\begin{array}{r} 4c_1 + 8c_2 + 4c_3 = 0 \\ 5c_1 + 8c_2 \\ - \quad - \\ \hline -c_1 + 4c_3 = 0 \end{array} \Rightarrow c_1 = -4c_3 \Rightarrow c_1 = 4c_3$$

Hence, $c_i \neq 0$, i = 1, 2, 3

The given set of vectors a_1, a_2, and a_3 are linearly dependent.

b) Define linear transformation and range of linear transformation. **(Unit-I, Q.No. 5, 6)**

Unit-II

3. a) In a class there are 10 boys and 5 girls. A committee of 4 students is to be selected from the class. Find the probability for the committee to contain at least 3 girls.

Ans:

The number of boys = n(B) = 10

The number of girls = n(G) = 5

Total number of students = n(S) = 15.

A committee has to select 4 students.

The probability that it contain at least 3 girls.

$= p(x \geq 3)$

$= p(x = 3 \text{ girls}) + P(x = 4 \text{ girls})$

$= \frac{10C_1 \times 5C_3}{15C_4} + \frac{10C_0 \times 5C_4}{15C_4}$

$= \frac{100}{1,365} + \frac{5}{1,365} = \frac{105}{1,365} = 0.0769$

---

b) A problem in statistics is given to the 3 students A, B, C whose chances of solving $\frac{1}{2}, \frac{1}{3}$ and $\frac{1}{4}$ respectively. What is the probability that the problem is solved?

Ans:

Let E_1, E_2, E_3 be the respective events of solving the problem and E_1, E_2, E_3 be the respective events of not solving the problem then.

$P(E_1) = \frac{1}{2}, P(E_2) = \frac{1}{3}, P(E_3) = \frac{1}{4}$

$\Rightarrow P(\bar{E}_1) = 1 - \frac{1}{2} = \frac{1}{2}$

$P(\bar{E}_2) = 1 - \frac{1}{3} = \frac{2}{3}$

$P(\bar{E}_3) = 1 - \frac{1}{4} = \frac{3}{4}$

$\therefore$ P(None solves the problem)

$= P(\bar{E}_1 \cap \bar{E}_2 \cap \bar{E}_3)$

$= P(\bar{E}_1) \cdot P(\bar{E}_2) \cdot P(\bar{E}_3)$ $[\because \bar{E}_1, \bar{E}_2, \bar{E}_3$ are independent]

$= \frac{1}{2} \times \frac{2}{3} \times \frac{3}{4} = \frac{1}{4}$

Hence P (the problem will be solved)

$= 1 - P$ (none solves the problem)

$= 1 - \frac{1}{4} = \frac{3}{4}$

(OR)

4. a) A manufacturer of cotter pins knows that 5% of his product is defective. Pins are sold in boxes of 100. He guarantees that not more than 10 pins will be defective. What is the approximate probability that a box will fail to meet the guaranteed quality?

Ans:

We are given n = 100

Let p = probability of a defective bulb = 5 % = 0.05

$\therefore$ m = mean number of defective bulbs in a box of 100 = np = 100 × 0.05 = 5

Since p is small , we can use poison's distribution.

Probability of x defective bulbs in a box of 100 is

$$P(X = x) = \frac{e^{-m}m^x}{x!} = \frac{e^{-5}5^x}{x!}, x = 0,1,2...$$

Probability that is box will fail to meet the guaranteed quality is

$$P(X > 10) = 1 - P(X \leq 10) = 1 - \sum_{x!}^{10} \frac{e^{-5}5^x}{x!} = 1 - e^{-5} \sum_{x=0}^{10} \frac{5^x}{x!}$$

b) Ten coins are thrown simultaneously. Find the probability of getting at least seven heads.

Ans:

The probability formula is defined as the possibility of an event to happen is equal to the ratio of the number of favorable outcomes and the total number of outcomes.

In the given case,

The probability of getting at least 7 when ten coins are tossed simultaneously is such that:

$$P(X \geq 7) = P(X = 7) + P(X = 9) + P(X = 10)$$
$$= (10C_3 + 10C_2 + 10C_1 + 1)\,(1/2)^{10}$$
$$= 176/(2^{10})$$
$$= 0.171875$$

Hence, when ten coins are tossed simultaneously the probability of getting at least 7 is 0.171875.

Unit - III

5. a) Define point estimation and interval estimation. **(Unit-III, Q.No. 10)**

b) What is the maximum error one can expect to make with probability 0.90 when using the mean of a random sample of size n = 64 to estimate the mean of population with $\sigma^2 = 2.56$.

Ans:

$$E = Z_{a/2} \cdot \frac{\sigma}{\sqrt{n}}$$

$$E = (1.645) \cdot \left(\frac{1.60}{\sqrt{64}}\right)$$

E = 0.329.

(OR)

6. a) Find 95% confidence limits for the mean of a normally distributed population from which the following sample was taken 15, 17, 10, 18, 16, 9, 7, 11, 16, 14.

Ans:

$c = 0.95$

$n = 10$

$$\bar{x} = \frac{15+17+10+18+16+9+7+11+16+14}{10}$$

$$= \frac{133}{10} = 13.3 = 13$$

$$s = \sqrt{\frac{(15-13)^2+(17-13)^2+(10-13)^2+(18-13)^2+(16-13)^2+(9-13)^2+(7+13)^2+(11-13)^2+(13-16)^2+(14-13)^2}{10-1}}$$

$$= \sqrt{\frac{4+16+9+25+9+16+36+4+9+1}{9}}$$

$$= \sqrt{\frac{129}{9}} = \sqrt{14.33} = 3.785$$

$df = n - 1 = 10 - 1 = 9$

$$\alpha = \frac{1-c}{2} = 0.025$$

$t_{a/2} = 2.26216$

The margin of error is :

$$E = t_{a/2} \times \frac{s}{\sqrt{n}} = 2.262 \times \frac{3.785}{\sqrt{10}} = 2.7076$$

The confidence interval:

$\bar{x} - E < \mu < \bar{x} + E$

$10 - 2.70 < \mu < 10 + 2.70$

$7.3 < \mu < 12.70$

b) A random sample of size 81 was taken whose variance is 20.25 and mean is 32, construct 98% confidence interval.

Ans:

Given $\bar{x}$ = Sample mean = 32, n = 81

$\sigma^2 = 20.25 \Rightarrow \sigma = 4.5$

and $z_{\alpha/2} = 2.33$ (For 98%)

We know that 98% confidence interval is $\left(\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$

Now $z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} = (2.33)\ \frac{4.5}{\sqrt{8.1}} = 1.165$

$\therefore$ Confidence interval = (32 – 1.29, 32 + 1.29) = (30.835, 33.165).

Unit - IV

7. a) A die is tossed 256 times and it turns up with an even digit 150 times. Is the die biased?

Ans:

Here n = 256, p = probability of getting an even digit (2 or 4 or 6) = $\frac{3}{6} = \frac{1}{2}$.

$q = 1 - p = \frac{1}{2}$ and $\mu = np = 256 \times \frac{1}{2} = 128$

$\sigma = \sqrt{npq} = \sqrt{(np)q} = \sqrt{128 \times 1/2} \equiv \sqrt{64} \equiv 8$

$\bar{x}$ = number of successes = 150

1. Null hypothesis H_0 : the die is unbiased
2. Alternate hypothesis H_1: the die is biased
3. level of significance, $\alpha = 0.05$
4. The test statistic = $Z = \frac{x - \mu}{\sigma} = \frac{150 - 128}{8} = \frac{22}{8} \equiv 2.75$

So $|Z| > 1.96$, the null hypothesis H_0 has to be rejected at 5% level of significance and we conclude that the die is unbiased, Hence the solution.

b) Write about (i) critical region and (ii) Two tailed test. **(Unit-IV, Q.No. 1, 8)**

(OR)

8. a) What is meant by level of significance?

Ans:

The level of significance is defined as the fixed probability of wrong elimination of null hypothesis when in fact, it is true. The level of significance is stated to be the probability of type I error and is preset by the researcher with the outcomes of error. The level of significance is the measurement of the statistical significance. It defines whether the null hypothesis is assumed to be accepted or rejected. It is expected to identify if the result is statistically significant for the null hypothesis to be false or rejected.

b) In a random sample of 160 workers exposed to a certain amount of radiation, 24 experienced some ill effects. Construct a 99% confidence interval for the corresponding true percentage.

Ans:

In a random sample of 160 workers exposed to a certain amount of radiation 24 experience some ill effects.

From the given dats n = 160 and x = 24

So the proportion value is given by,

$$\frac{x}{n} = \frac{24}{160} = 0.15$$

Here we have to construct the 99% confidence interval for the true proportion so $\alpha = 0.01$.

For the large sample the $(1 - \alpha)100\%$ confidence interval is given as follows:

$$\frac{x}{n} - z_{a/2}\sqrt{\frac{\frac{x}{n}\left(1-\frac{x}{n}\right)}{n}} < p < \frac{x}{n} + z_{\alpha/2}\sqrt{\frac{\frac{x}{n}\left(1-\frac{x}{n}\right)}{n}}$$

Where

$$Z_{a/2} = Z_{0.01/2}$$
$$= 2.58$$

By substituting the values we get,

$$0.15 - 2.58\sqrt{\frac{0.15\ (1-0.15)}{160}} < p < 0.15 + 2.58\sqrt{\frac{0.15(1-0.15)}{160}}$$

$$0.0772 < p < 0.2228$$

Therefore the 99% confidence interval for the true proportion is (0.0772.0.2228).

Unit-V

9. Four methods are under development for making discs of a super conducting material. Fifty discs are mad by each method and they are checked for super conductivity when cooled with liquid.

	1st method	2nd method	3rd method	4th method
Super conductors	31	42	22	25
Failures	19	8	28	25

Test the significance difference between the proportions of super conductors at 0.05 level.

Ans:

1. Null Hypothesis H_0: $p_1 = p_2 = p_3 = p_4$
2. Alternative Hypothesis H_1 : $p_1 \neq p_2 \neq p_3 \neq p_4$
3. Level of significance, $\alpha = 0.05$
4. Computations.

	Ist Method	**2nd Method**	**3rd Method**	**4th Metod**	**Total**
Super Conductors	31	42	22	25	120
Failures	19	8	28	25	80
Total	50	50	50	50	200

Table of expected frequencies

$\frac{50\times120}{200} = 30$	$\frac{50\times120}{200} = 30$	$\frac{50\times120}{200} = 30$	$\frac{50\times120}{200} = 30$	120
$\frac{50\times80}{200} = 20$	$\frac{50\times80}{200} = 20$	$\frac{50\times80}{200} = 20$	$\frac{50\times80}{200} = 20$	80
50	50	50	50	200

Calculation for χ^2

Observed	Expected Frequency (O_i)	$(O_i - E_i)^2$ Frequency (E_i)	$\frac{(O_i - E_i)^2}{E_i}$
31	30	1	1/30
42	30	144	144/30
22	30	64	64/30
25	30	25	25/30
19	20	1	1/20
8	20	144	144/20
28	20	64	64/20
25	20	25	25/20
			19.5

$\therefore$ Calculated $\chi^2 = 19.5$

Tabulated χ^2 for $(4 - 1)(4 - 1) = 9$ d.f. at 5% level of significance is 7.815

Since calculated χ^2 > tabulated χ^2, we reject the null hypothesis H_0 i.e., There is a significant difference between the proportions.

(OR)

10. Given the bi-variate data

X	1	5	3	2	1	1	7	3
Y	6	1	0	0	1	2	1	5

a) Find the regression line of Y on X and hence predict Y if X = 10.

b) Fit a Regression line of X on Y and hence predict X if Y = 2.5.

Ans:

a) Find the regression line of Y on X and hence predict Y if X = 10.

The regression line of Y on X is as follow:

$Y = \beta_0 + \beta_1 x$, where the β_0 and β_1 are the intercept and slope of the line and need to be estimated.

Table for calculation

X	Y	XY	X^2
1	6	6	1
5	1	5	25
3	0	0	9
2	0	0	4
1	1	1	1
1	2	2	1
7	1	7	49
3	5	15	9
23	16	36	99

$$\beta_1 = \frac{n\Sigma xy - \Sigma cx\ \Sigma y}{n\Sigma x^2 - (\Sigma x)^2} = \frac{8\times 36 - 23\times 16}{8\times 90 - 23^2} = -0.3042$$

$$\beta_1 = \frac{\Sigma y - \beta_1 \Sigma x}{n} = \frac{16 - (-0.3042)\times 23}{8} = 2.8745$$

Therefore, the regression line of Y on X is as follows:

$Y = 2.8745 - 0.3042X$

Now, the predicted value of Y if X = 10 is as follows:

$Y = 2.8745 - 0.3042 \times 10$

$= -0.1675$

Thus, the predicted value of Y if X = 10 is – 0.1675.

b) Fit a Regression line of X on Y and hence predict X if Y = 2.5.

$X = \beta_0 + \beta_1 Y$, where the β_0 and β_1 are the intercept and slope of the line and need to be estimated.

Table for calculation

X	Y	XY	X^2
1	6	6	36
5	1	5	1
3	0	0	0
2	0	0	0
1	1	1	1
1	2	2	4
7	1	7	1
3	5	15	25
23	16	36	68

$$\beta_1 = \frac{n\Sigma xy - \Sigma x\ \Sigma y}{n\Sigma Y^2 - (\Sigma Y)^2}$$

$$= \frac{8\times 36 - 23\times 16}{8\times 68 - 16^2} = \frac{80}{288} = -0.2778$$

$$\beta_1 = \frac{\Sigma x - \beta_1 \Sigma y}{n} = \frac{23 - (-0.2778)\times 16}{8} = 3.4306$$

Therefore, the regression line of Y on X is as follows:

$X = 3.4306 - 0.2778 \times 2.5$

$= 2.7361$

Now, the predicted value of X if Y = 2.5 is as follows:

$Y = 3.4306 - 0.2778 \times 2.5$

$= 2.7361$

Thus, the predicted value of X if Y = 2.5 is 2.7361.