STATISTICAL SOFTWARE USING R

PRACTICAL-I (Papers on 101ST24 and 102ST24)

M.Sc., STATISTICS First Year

Semester – I, Paper-V

Lesson Writers

Dr. S. Bhanu Prakash
Assistant Professor
Freshman Engineering Department
Dept. of Mathematics & Statistics
Godavari Global University
Rajamahendravaram

Editor:
Dr. R. Vishnu Vardhan
Associate Professor
Department of Statistics
Pondicherry University
.

Director, I/c
Prof. V.VENKATESWARLU
MA.,M.P.S., M.S.W.,M.Phil., Ph.D.

CENTREFORDISTANCEEDUCATION
ACHARAYANAGARJUNAUNIVERSITY
NAGARJUNANAGAR – 522510
Ph:0863-2346222,2346208,
0863-2346259(Study Material)
Website: www.anucde.info
e-mail:anucdedirector@gmail.com

M.Sc., STATISTICS - STATISTICAL SOFTWARE USING R

First Edition2025

No. of Copies :

Published by:

Prof. V.VENKATESWARLU,

Director, I/C

Centre for Distance Education,

Acharya Nagarjuna University

Printed at:

FOREWORD

Since its establishment in 1976, Acharya Nagarjuna University has been forging ahead in the path of progress and dynamism, offering a variety of courses and research contributions. I am extremely happy that by gaining 'A^+' grade from the NAAC in the year 2024, Acharya Nagarjuna University is offering educational opportunities at the UG, PG levels apart from research degrees to students from over 221 affiliated colleges spread over the two districts of Guntur and Prakasam.

The University has also started the Centre for Distance Education in 2003-04 with the aim of taking higher education to the doorstep of all the sectors of the society. The centre will be a great help to those who cannot join in colleges, those who cannot afford the exorbitant fees as regular students, and even to housewives desirous of pursuing higher studies. Acharya Nagarjuna University has started offering B.Sc., B.A., B.B.A., and B.Com courses at the Degree level and M.A., M.Com., M.Sc., M.B.A., and L.L.M., courses at the PG level from the academic year 2003-2004 onwards.

To facilitate easier understanding by students studying through the distance mode, these self-instruction materials have been prepared by eminent and experienced teachers. The lessons have been drafted with great care and expertise in the stipulated time by these teachers. Constructive ideas and scholarly suggestions are welcome from students and teachers involved respectively. Such ideas will be incorporated for the greater efficacy of this distance mode of education. For clarification of doubts and feedback, weekly classes and contact classes will be arranged at the UG and PG levels respectively.

It is my aim that students getting higher education through the Centre for Distance Education should improve their qualification, have better employment opportunities and in turn be part of country's progress. It is my fond desire that in the years to come, the Centre for Distance Education will go from strength to strength in the form of new courses and by catering to larger number of people. My congratulations to all the Directors, Academic Coordinators, Editors and Lesson-writers of the Centre who have helped in these endeavors.

Prof. K. Gangadhara Rao
M.Tech.,Ph.D.,
Vice-Chancellor I/c
Acharya Nagarjuna University

M.Sc.–Statistics

Syllabus SEMESTER-I

105ST24: STATISTICAL SOFTWARE USING R

(Papers on 101ST24 and 102ST24)

101ST24 – PROBABILITY THEORY AND DISTRIBUTIONS

Practical 1: Fitting of binomial distribution and test for goodness of fit

1. Write R-Code for fitting of binomial distribution and test for goodness of f it. R-code for Execute your R-code for the following data.

A die is thrown 60 times with the following results.

Face	1	2	3	4	5	6
Frequency	8	7	12	8	14	11

Practical 2: Fitting of poisson distribution and test for goodness of fit

2. The following table gives the count of yeast cells in square of a cyclometer. A square millimeter is divided into 400 equal squares and the number of these squares containing 0,1,2,... cells are recorded.

Number of cells	0	1	2	3	4	5	6	7	8	9	10
Frequency	2	18	43	53	86	70	54	37	18	10	5
Number of cells	11	12	13	14	15	16					
Frequency	2	2	0	0	0	0					

Fit a poisson distribution to the data and test the good ness of fit.

Practical 3: Fitting of normal distribution and test for goodness of fit

3. Write the necessary R code for solving the following problem and execute the same on the system. Fit normal distribution to the following data and test for goodness of fit.

Marks(X)	No.of Students
15-19	9
20-24	11
25-29	10
30-34	44
35-39	45
40-44	54
45-49	37
50-54	36
55-59	8
60-64	5
65-69	1

Practical4: Fitting of logistic distribution and test for goodness of fit

4. Write the necessary R code for solving the following problem and execute the same on the system. Fit a logistic distribution to the following data and test for goodness of fit.

Class Interval	**Frequency**
11-13	08
13-15	24
15-17	42
17-19	65
19-21	36
21-23	16
23-25	09

Practical 5: Fitting of exponential distribution and test for goodness of fit

5. Write the necessary R code for solving the following problem and execute the same on the system.

The distribution of age at the marriage of grooms with brides of the following groups:

Age Groups	15-19	19-23	23-27	27-31	31-35
No. of Groups	08	25	42	18	07

Fit exponential distribution for the given data and also test whether the fit is good fit or not.

Practical 6: Practical based on application of Multi nomial Distribution.

6. A company is conducting a survey with **100 respondents,** and they are asked to choose one of 3 **options** for a new product (let's call them **Option A, Option B,** and **Option C).**The probabilities of each option being selected are:

- **OptionA:**0.4
- **OptionB:**0.35
- **OptionC:**0.25

The company wants to model the number of respondents choosing each option in **100 trials.**

Practical 7: **Practical based on Conditional Probability**

7. In a deck of **52 playing cards,** what is the probability of drawing a **heart** given that the card drawn is a **red card?** (Recall that there are 26 red cards, and half of them are hearts).

Practical 8:**Practical based on Geometrical Probability**

8. Buffon's Needle is a famous problem in probability. You drop a needle of length I on to a floor with parallel lines spaced **d** apart. The goal is to estimate the probability that the needle will cross one of the lines.

For simplicity, let the needle length I=**d,** which is the simplest case. The probability of Crossing a line is given by the formula **P=***1rJ2.*

We'll simulate this experiment and estimate the probability by dropping a needle many times.

102ST24-STATISTICALCOMPUTINGUSINGR

Practical 1:Computing of Mean, Median, Geometric Mean and SD

Problem: Write a R code for computing Mean, Median, Geo metric Mean and SD for the frequency distribution.
Execute R-code for the following data

Marks	15-19	20-24	25-29	30-34	35-39	40-44	45-49	50-54	55-59	60-64	65-69
No.of Students	9	11	10	44	45	54	37	36	18	5	I

Practical 2:One Samplet-Test

Problem: A random sample of IO students had the following I.Q. 70, 12 110,101, 885,95,9 10100.
Do this data support the assumptions of a population mean I.Q. of 100?

Practical3:TwoSamplet-Test

Problem: Below are given the gainin weights (in kgs) of pigs fat of two digits A and B Diet A

Diet A	25	32	30	34	24	14	32	24	30	31	35	25	-	-	-
Diet B	44	34	22	10	47	31	40	30	32	35	18	21	35	29	22

Practical 4: Newton - Raphs on Method

Problem: Solve the equation $x^3-2x-5=0$ using Newton-Raphs on method by writing necessary R-code and executing the same.

Practical 5: Completely Randomized Design (CRD)

Problem: The weights in gm of a number of copper wires, each of length I meter,were obtained. These are classified according to the die from which they come.

Die No.				
I	II	III	**IV**	V
1.30	1.28	1.32	1.31	1.30
1.32	1.35	1.29	1.29	1.32
1.36	1.33	1.31	1.33	1.30
1.35	1.34	1.28	1.31	1.33
1.32	NA	1.33	1.32	NA
1.37	NA	1.30	NA	NA

Setup an analysis of variance table to test the significance of the difference between the weights due to different dies. Compute CRD analysis in R.

Practical 6: **Randomized Block Design (RBD)**
Problem: Setup a table of analysis of variance for yields of three strains of wheat planted in five randomized blocks.

Strains	Blocks				
A	20	21	23	16	20
B	18	20	17	15	25
C	25	28	22	28	32

Write the necessary **R** code for RBD analysis

Practical 7: **Fitting Regression Lines**

Problem: A group of students recorded their study hours and corresponding exam scores:

Study Hours: 2, 3, 5, 7, 8,I0, 12, 14

Exam Scores: 50,55,65,70,75,80,90,95

Analyze the relationship between study hours and exams cores by fitting a regression line.

Practical 8: High-Level Plotting-Creating Different Types of Plots

Problem

Consider a data set containing the monthly sales (in thousands)of a product in different regions:

Practical 9: Low-Level Plotting-Adding Elements to a Plot
Problem: For the following data set:

X values: 1,2,3,4,5,6,7,8

Y values: 2,4,6,8,10,12,14,16

Create a scatter plot and customize it by adding:

- A horizontal reference line at Y=10.
- A vertical reference line at *X=S.*
- Gridlines and custom points.

Practical10:Q-Q Plot and Pie Chart

Problem: Given the following data sets:

I. Normal data: X={2,4,6,8,10,12,14,16}.

2.Categories:**A**(25%),B(35%),C(20%),D(20%).

Create:

1. A Q-Q plot to check normality for X.
2. A pie chart to visualize the categorical proportions.

CONTENTS

Course-105ST24: PRACTICAL-I Probability Theory and Distribution -------Lab

LAB Exercise1:

Fitting of binomial distribution and test for goodness of fit

Problem: Write R-Code for fitting of binomial distribution and test for goodness of fit. Execute your R-code for the following data.

A die is thrown 60 times with the following results.

Face	1	2	3	4	5	6
Frequency	8	7	12	8	14	11

Aim: Fitting of Binomial Distribution for the given data and test for goodness of fit.

Procedure: We have to estimate 'P' value from the given data using the following mean formula.

$$np = \bar{X} = \sum fixi / N$$

$$\hat{P} = \left(\sum fixi / N\right) / N$$

Where N= $\sum fi$

N is number of trails

We have to find binomial probabilities using the probability mass function .

$$b\left(p(x)\right) = n_{C_x} p^x q^{n-x} \quad x=0, 1, 2\ldots\ldots n$$

Now, we have to find the expected frequencies using the formula.

$$e_x = Nbp(x)$$

Test for Goodness of Fit:

We have to test the goodness of fit using the χ^2 formula

$$\chi^2 = \sum\left[(o_i - e_i)^2 / e_i\right] \sim \chi^2_{(n)} df$$

The critical chi-square value $\chi^2_{(0.05)}$ at 5% los

Conclusion: If calculated χ^2 value is less than critical χ^2 value (0.05) df. We accept the null hypothesis at 5% los which means the binomial distribution is fitted well to the given data, otherwise if χ^2 greater than $\chi^2_{(n)}$ then the binomial distribution is not well fitted for the data.

R code:

```
cat("\n Enter observed Frequencies:")
f=scan();
n=length(f)-1;x=0:n
N=sum(f)
#'p' value is given in the problem
#p=0.5
#calculating 'p' value from the given data
mean=sum(f*x)/N;
p=mean/n
cat("\n Estimate Value of Probability of Success'p'(from the data)=",p);
#COMPUTES EXPECTED FREQUENCIES
#bprob=dbinom(x,n,p);
bprob=choose(n,x)*p^x*(1-p)^(n-x);
 cat("\n check on sum of probabilities=",sum(bprob));
 e=N*bprob;
cat("\n Observed frequencies:",f); cat("\n Expected
frequencies:",round(e,0));
 cat("\n Sum of observed frequencies=",sum(f));
 cat("\n Sum of expected frequencies=",sum(e));
#TEST FOR GOODNESS OF FIT USING CHI-SQUARE TEST
CHSV=0;
CHSV=sum((f-e)^2/e);
 cat("\n CALCULATED chi-square value=",CHSV);
 CRCV=qchisq(0.95,n); cat("\n critical chi-square value=",CRCV);
cat("(df=",n,")");
 if(CHSV<=CRCV)
cat("\n\n binomial distribution is WELL FITTED to the given data")
if(CHSV>CRCV)cat("\n\n Binomial distribution is NOT FITTED to the given data")
cat("\n***End of computation***\n")
```

Output :

```
> source("C:\\Users\\ADMIN\\Desktop\\binom.R")
```

Enter observed Frequencies:1: 8

2: 7

3: 12

4: 8

5: 14

6: 11

7:

Read 6 items

Estimate Value of Probability of Success'p'(from the data)= 0.5533333

check on sum of probabilities= 1

Observed frequencies: 8 7 12 8 14 11

Expected frequencies: 1 7 16 20 13 3

Sum of observed frequencies= 60

Sum of expected frequencies= 60

CALCULATED chi-square value= 73.84244

critical chi-square value= 11.0705(df= 5)

Binomial distribution is NOT FITTED to the given data

End of computation

Inference:

The calculated χ^2 value is greater than critical χ^2 value (0.05) df. We reject the null hypothesis at 5% los which means the binomial distribution is not well fitted to the given data.

LAB Exercise2:

Fitting of poisson distribution and test for goodness of fit

Problem: The following table gives the count of yeast cells in square of a cyclometer. A square millimeter is divided into 400 equal squares and the number of these squares containing 0,1,2,…cells are recorded.

Number of cells	0	1	2	3	4	5	6	7	8	9	10
Frequency	2	18	43	53	86	70	54	37	18	10	5
Number of cells	11	12	13	14	15	16					
Frequency	2	2	0	0	0	0					

Fit a poisson distribution to the data and test the goodness of fit.

Aim: Fitting of Poisson distribution for the given data and test for goodness of fit.

Procedure: The estimated of poisson distribution population mean λ is given by

$\lambda = \bar{x} = \sum fixi / N$ x=0,1,2……n

Where $N = \sum fi$

We have to find poisson probabilities using the probability mass function

$p(x) = e^{-\lambda}\lambda^{x} / x!$, x=0,1,….n

Where 'n' is the number of classes

Now, we have to find the expected frequencies using the formula $e_x = Np(x)$

Test for goodness of fit: We have to test the goodness of fit using the χ^2 formula

$$\chi^2 = \sum \left[(o_i - e_i)^2 / e_i\right] \square \chi^2_{(n)} df$$

The critical chi-square value $\chi^2_{(0.05)}$ at 5% los

Conclusion: If calculated χ^2 value is less than critical χ^2 value (0.05) df. We accept the null hypothesis at 5% los which means the poisson distribution is fitted well to the given data, otherwise if χ^2 greater than $\chi^2_{(n)}$ then the poisson distribution is not well fitted for the data.

R code:

```
cat("\nEnter observed Frequencies:");
f=scan()
n=length(f);
```

```
n=n-1;
x=0:n
N=sum(f)
#poisson mean value is given in the problem
#mean=0.5
#calculating 'mean' value from the given data
mean=sum(f*x)/N
cat("\n Estimated value of poission mean value (from the  data)=",mean);
#COMPUTES EXPECTED FREQUENCIES
pprob=dpois(x,mean);
cat("\n checkon sum of probabilities=",sum(pprob));
e=N*pprob;  cat("\n observed frequencies:",f);
cat("\n Expected frequencies:",round(e,0));
cat("\n sum of observed frequencies=",sum(f));
cat("\n sum of expected frequencies=",sum(e));
#TEST FOR GOODNESS OF FIT USING CHI-SQUARE TEST
CHSV=0;
CHSV=sum((f-e)^2/e);
cat("\n CALCULATED chi-square value=",CHSV);
CRCV=qchisq(0.95,n);  cat("\n critical chi-square value",CRCV);
cat("(d.f.=",n,")");
if (CHSV<=CRCV)
cat("\n\n poisson distribution is WELL FITTED to the given data");
if (CHSV>CRCV)
cat("\n\n poisson distribution is NOT FITTED to the given data");
cat("\n ***End of computation ***\n");
```

Output :

```
> source("C:\\Users\\ADMIN\\Desktop\\poisson.R")

Enter observed Frequencies:1: 2
2: 18
3: 43
```

4: 53

5: 86

6: 70

7: 54

8: 37

9: 18

10: 10

11: 5

12: 2

13: 2

14: 0

15: 0

16: 0

17: 0

18:

Read 17 items

Estimated value of poission mean value (from the data)= 4.675

checkon sum of probabilities= 0.9999914

observed frequencies: 2 18 43 53 86 70 54 37 18 10 5 2 2 0 0 0 0

Expected frequencies: 4 17 41 64 74 69 54 36 21 11 5 2 1 0 0 0 0

sum of observed frequencies= 400

sum of expected frequencies= 399.9966

CALCULATED chi-square value= 7.145309

critical chi-square value 26.29623(d.f.= 16)

poisson distribution is WELL FITTED to the given data

***End of computation ***

Inference :

The calculated χ^2 value is less than critical χ^2 value (0.05) df. We accept the null hypothesis at 5% los which means the poisson distribution is well fitted to the given data.

LAB Exercise 3:

Fitting of Exponential distribution and test for goodness of fit

Problem: Write the necessary R code for solving the following problem and execute the same on the system.

The distribution of age at the marriage of grooms with brides of the following groups:

Age Groups	15-19	19-23	23-27	27-31	31-35
No. of Groups	08	25	42	18	07

Fit exponential distribution for the given data and also test whether the fit is good fit or not.

Aim: Fitting of Exponential Distribution for the given data and test for goodness of fit.

Procedure: We have to estimate λ value from the given data using the following mean formula.

$1/\lambda = \bar{x} = \sum fixi / N$

$\hat{\lambda} = 1/\sum fixi / N$

Where $N = \sum fi$

We have to find exponential probabilities.

The pdf

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

Now, we have to the expected frequencies using the formula $e^{|}_{x} = N^{|} f(x)$

Test for goodness of fit: We have to test the goodness of fit using the χ^2 formula

$$\chi^2 = \sum \left[(o_i - e_i)^2 / e_i \right] \square \chi^2_{(n-1)} df$$

The critical chi-square value $\chi^2_{(0.05)}$ at 5% los

Conclusion: If calculated χ^2 value is less than critical χ^2 value (0.05) df. We accept the null hypothesis at 5% los which means the exponential distribution is fitted well to the given data; otherwise if χ^2 greater than $\chi^2_{(n-1)}$ then the exponential distribution is not well fitted for the data.

R code:

```
cat("\n Enter Class Mid values:"); X=scan();
g=length(X);
cat("\n Enter frequency data=");
F=scan();
N=sum(F);
mean=sum(F*X)/N;
rate=1/mean;
cat("\n Estimate of Rate(parameter)of exponential
distribution=",round(rate, digit=4));
cw=X[2]-X[1];
cw=cw/2;
cat("\n Class Intervals:",paste(X-cw, X+cw));
P=pexp(X+cw, rate)- pexp(X-cw, rate);
P[1]=pexp(X[1]+cw, rate);
P[g]=1-pexp(X[g]-cw, rate);
E=round(N*P);
cat("\n Observed Frequencies:",F);
cat("\n Expected Frequencies:",E);
cat("\n TEST FOR GOODNESS OF FIT USING USER DEFINED CHISQUARE
TEST:\n");
CHSV=sum((F-E)^2/E);
cat("\n Computed Chi^2 value=",round (CHSV, digit=4));
CRCV=qchisq(0.95,g);
cat("\n Critical Chi^2 value(at 5%los with df:",g-1,")=", round(CRCV, digits=4));
cat("\n CONCLUSION");
if(CHSV<=CRCV)
cat("\n EXPONENTIAL distribution is WELL FITTED for the given data");
if(CHSV>CRCV)
cat("EXPONENTIAL distribution is NOT WELL FITTED for the given data");
chit=chisq.test (F, p=E, rescale.p=T)
print(chit);
cat("\n***End of computation***\n");
```

Output :

```
> source("C:\\Users\\ADMIN\\Desktop\\exp.R")
Enter Class Mid values:1: 17
2: 21
3: 25
4: 29
5: 33
6:
Read 5 items
Enter frequency data=1: 8
2: 25
3: 42
4: 18
5: 7
6:
Read 5 items
Estimate of Rate(parameter)of exponential distribution= 0.0406
Class Intervals: 15 19 19 23 23 27 27 31 31 35
Observed Frequencies: 8 25 42 18 7
Expected Frequencies: 54 7 6 5 28
TEST FOR GOODNESS OF FIT USING USER DEFINED CHISQUARE TEST:

Computed Chi^2 value= 351.0209
Critical Chi^2 value(at 5%los with df: 4 )= 11.0705
CONCLUSION
EXPONENTIAL distribution is NOT WELL FITTED for the given data
Chi-squared test for given probabilities

data:  F
X-squared = 351.02, df = 4, p-value < 2.2e-16
 ***End of computation***
```

Inference :

The calculated χ^2 value is greater than critical χ^2 value (0.05) df. We reject the null hypothesis at 5% los which means the exponential distribution is not well fitted to the given data.

LAB Exercise 4:

Fitting of Logistic distribution and test for goodness of fit

---

Problem: Write the necessary R code for solving the following problem and execute the same on the system.Fit a logistic distribution to the following data and test for goodness of fit.

Class Interval	Frequency
11-13	08
13-15	24
15-17	42
17-19	65
19-21	36
21-23	16
23-25	09

Fitting of logistic distribution and test for goodness of fit.

Aim: To fit a logistic distribution to the given data and obtain expected logistic frequencies and also to test for goodness of fit.

Procedure:

To estimate of logistic distribution are given by $\mu = \sum fixi / N$

$SD = \sqrt{\sum fixi^2 / N - \mu^2}$ **:** $\sigma = SD\sqrt{3} / \pi$

We have to find the expected logistic frequencies using the formula

$E = NP_i$

Where N is total frequencies $P_i = F(x_{i+1}) - F(x_i); i = 1, 2.......g$

$$F(x_i) = 1/1 + \exp\left[\frac{-(x_i - \mu)}{\sigma}\right]$$

In particular $P_1 = F(x_2)$

$$P_g = 1 - F(x_g)$$

Test for goodness of fit: We have to test the goodness of fit using the χ^2 formula

$$\chi^2 = \sum \left[(o_i - e_i)^2 / e_i\right] \sim \chi^2_{(g-1)} df$$

Where (g-1) is df

The critical chi-square value $\chi^2_{(0.05)}$ at 5% los

Conclusion: If calculated χ^2 value is less than critical χ^2 value (0.05) df. We accept the null hypothesis at 5% los which means the logistic distribution is fitted well to the given data, otherwise if χ^2 greater than $\chi^2_{(g-1)}$ then the logistic distribution is not well fitted for the data.

R-CODE

```
cat("\n Enter Class mid values:");
X=scan();
g=length(X);
cat("\n Enter frequency data=");
F=scan();
N=sum(F);
mu=sum(F*X)/N ;
sd=sqrt(sum(F*(X-mu)^2)/(N-1)) ;
sigma=sd*sqrt(3)/pi;
cat("\n Estimates of logistic parameters: mu=",round(mu,digit=4));
cat("sigma=",round(sigma, digit=4));
cat("\n Observed Frequencies:",F,"\n");
cw=X[2]-X[1];
cw=cw/2;
P=plogis(X+cw, mu, sigma)-plogis(X-cw, mu, sigma);
P[1]=plogis(X[1]+cw, mu, sigma);
P[g]=1-plogis(X[g]-cw, mu, sigma);
E=round(N*P);
cat("\n Expected Frequencies:",E,"\n");
cat("\n TEST FOR GOODNESS OF FIT USING OUR OWN CHI-SQUARE TEST:");
CHSV=sum((F-E)^2/E);
cat("\n Computed Chisquare Value:",round(CHSV,digit=4));
CRCV=qchisq(0.95,g-1);
cat("\nCritical Chi-square value:",round(CRCV,digits=4));
cat("\nCONCLUSION:");
```

```
if(CHSV<=CRCV)
cat("\n Logistic distribution is WELL FITTED for the given data.")
if (CHSV>CRCV)
cat("\n Logistic distribution is NOT FITTED for the given data.");
cat("\n\n Goodness of fit test using built-in Chi-square test:");
chit=chisq.test(F,p=E,rescale.p=T);
print(chit);
cat("\n***End of computation***")
```

Output :

```
> source("C:\\Users\\ADMIN\\Desktop\\logis.R")

 Enter Class mid values:1: 12
2: 14
3: 16
4: 18
5: 20
6: 22
7: 24
8:
Read 7 items

 Enter frequency data=1: 8
2: 24
3: 42
4: 65
5: 36
6: 16
7: 9
8:
Read 7 items
Estimates of logistic parameters: mu= 17.81sigma= 1.542
Observed Frequencies: 8 24 42 65 36 16 9
```

Expected Frequencies: 8 19 46 62 41 16 7

TEST FOR GOODNESS OF FIT USING OUR OWN CHI-SQUARE TEST:

Computed Chisquare Value: 2.99

Critical Chi-square value: 12.5916

CONCLUSION:

Logistic distribution is WELL FITTED for the given data.

Goodness of fit test using built-in Chi-square test:

Chi-squared test for given probabilities

data: F

X-squared = 2.97, df = 6, p-value = 0.8126

End of computation

Inference :

The calculated χ^2 value is less than critical χ^2 value (0.05) df. We accept the null hypothesis at 5% los which means the logistic distribution is well fitted to the given data.

LAB Exercise 5:

Fitting of Normal Distribution and test for goodness of fit :

Problem: Write the necessary R code for solving the following problem and execute the same on the system. Fit normal distribution to the following data and test for goodness of fit.

Marks (X)	No. of Students
15-19	9
20-24	11
25-29	10
30-34	44
35-39	45
40-44	54
45-49	37
50-54	36
55-59	8
60-64	5
65-69	1

Aim: Fitting of Normal Distribution for the given data and test for goodness of fit.

Procedure: The estimates of the normal distribution is given by

$$\mu = \sum \frac{fx}{N}$$

$$\sigma = SD = \sqrt{\frac{\sum fx^2}{N} - \mu^2}$$

We have to find the expected normal frequencies using the formula

$E = NP_i$ Where N is total frequencies

$$P_i = F(x_{i+1}) - F(x_i); i = 1, 2.......g$$

$$F(x_i) = \frac{1}{\sigma\sqrt{2\pi}} e^{-1/2\left(\frac{x-\mu}{\sigma}\right)^2}$$

In particular $P_1 = F(x_2)$

$$P_g = 1 - F(x_g)$$

Now, we have to find the expected frequencies using the formula

$$e_x = N * P$$

Test for goodness of fit: We have to test the goodness of fit using the χ^2 formula

$$\chi^2 = \sum \left[(o_i - e_i)^2 / e_i\right] \sim \chi^2_{(g-1)} df$$

Where (g-1) is df

The critical chi-square value $\chi^2_{(0.05)}$ at 5% los

Conclusion: If calculated χ^2 value is less than critical χ^2 value (0.05) df. We accept the null hypothesis at 5% los which means the normal distribution is fitted well to the given data, otherwise if χ^2 greater than $\chi^2_{(g-1)}$ then the normal distribution is not well fitted for the data.

R-CODE

```
cat("\n Enter Class mid values:");
X=scan();
g=length(X)
cat("\n Enter frequency data=");
F=scan();
N=sum(F);
mu=sum(F*X)/N
var=sum(F*(X-mu)^2)/(N-1)
sigma=sqrt(var)
cat("\n  Estimates  of  normal  parameters:  mu=",  round(mu,digit=4));
cat("sigma=",round(sigma,digit=4));
cat("\n Observed  Frequencies:",F,"\n");
cw=X[2]-X[1];
cw=cw/2;
P=pnorm(X+cw, mu,sigma)-pnorm(X-cw, mu, sigma);
P[1]=pnorm(X[1]+cw, mu, sigma);
P[g]=1-pnorm(X[g]-cw, mu,sigma);
E=round(N*P);
cat("\n Expected Frequencies:",E,"\n");
cat("\n TEST FOR GOODNESS OF FIT USING OUR OWN CHI-SQUARE TEST:");
 CHSV=sum((F-E)^2/E);
cat("\n Computed Chi-square Value:", round(CHSV,digit=4));
CRCV=qchisq(0.95,g-1);
cat("\nCritical Chi-square value:", round(CRCV,digits=4));
cat("\nCONCLUSION:");
```

```
if(CHSV<=CRCV)
cat("\n Normal distribution is WELL FITTED for the given data.");
if(CHSV>CRCV)
cat("\n Normal distribution is NOT FITTED for the given data.")
cat("\n\n Goodness of fit test using built-in Chisquare test:");
chit=chisq.test(F,p=E,rescale.p=T);
print(chit);
cat("***End of computation***\n");
```

Output :

```
source("C:\\Users\\ADMIN\\Desktop\\normal.R")

Enter Class mid values:1: 17
2: 22
3: 27
4: 32
5: 37
6: 42
7: 47
8: 52
9: 57
10: 62
11: 67
12:
Read 11 items
Enter frequency data=1: 9
2: 11
3: 10
4: 44
5: 45
6: 54
7: 37
8: 36
9: 8
```

10: 5

11: 1

12:

Read 11 items

Estimates of normal parameters: mu= 40.1923sigma= 10.0001

Observed Frequencies: 9 11 10 44 45 54 37 36 8 5 1

Expected Frequencies: 5 10 22 37 49 51 41 26 13 5 2

TEST FOR GOODNESS OF FIT USING OUR OWN CHI-SQUARE TEST:

Computed Chi-square Value: 18.3323

Critical Chi-square value: 18.307

CONCLUSION:

Normal distribution is NOT FITTED for the given data.

Goodness of fit test using built-in Chisquare test:

Chi-squared test for given probabilities

data: F

X-squared = 18.399, df = 10, p-value = 0.0486

End of computation

Inference :

The calculated χ^2 value is greater than critical χ^2 value (0.05) df. We reject the null hypothesis at 5% los which means the normal distribution is not well fitted to the given data.

LAB Exercise 6:

Practical based on application of Multinomial Distribution.

Problem: A company is conducting a survey with **100 respondents**, and they are asked to choose one of **3 options** for a new product (let's call them **Option A**, **Option B**, and **Option C**). The probabilities of each option being selected are:

- **Option A**: 0.4
- **Option B**: 0.35
- **Option C**: 0.25

The company wants to model the number of respondents choosing each option in **100 trials**.

Aim: To model the number of respondents choosing each option in a survey using Multinomial Distribution.

Procedure:

1. Set the parameters for the number of trials (n_trials) and the probabilities of selecting each option (probabilities).
2. Use the rmultinom() function to simulate the multinomial distribution with the given parameters.
3. Set a seed value for reproducibility.
4. Convert the result into a data frame to represent the number of respondents choosing each option.
5. Display the result.

R- code

```
# Set the parameters
n_trials <- 100
probabilities <- c(0.4, 0.35, 0.25)

# Simulate the multinomial distribution
set.seed(123) # For reproducibility
result <- rmultinom(1, size = n_trials, prob = probabilities)
```

```
# Display the result
result_df <- data.frame(Option_A = result[1,], Option_B = result[2,], Option_C = result[3,])
print(result_df)
```

Output :

```
> source("C:\\Users\\ADMIN\\Desktop\\multino.R")
  Option_A  Option_B  Option_C
1       39        36        25
```

Inference:

The simulation provides the expected number of respondents selecting each option (A, B, and C) based on the specified probabilities. This helps the company understand the likely distribution of choices among respondents.

LAB Exercise 7:

Practical based on Conditional Probability

Problem: In a deck of 52 playing cards, what is the probability of drawing a heart given that the card drawn is a red card? (Recall that there are 26 red cards, and half of them are hearts).

Aim: To calculate the conditional probability of drawing a heart given that the card drawn is a red card in a deck of 52 playing cards.

Procedure (Conditional Probability):

1. Identify the number of red cards and hearts in a standard deck of 52 playing cards.
2. Recognize that there are 26 red cards, out of which 13 are hearts.
3. Use the formula for conditional probability:
   P(Heart | Red) = P(Heart and Red) / P(Red)
4. Substitute the values: P(Heart | Red) = 13 / 26
5. Simplify to get the result.

R Code :

```
# Conditional Probability Calculation
red_cards <- 26
hearts <- 13
prob_heart_given_red <- hearts / red_cards
cat("Probability of drawing a heart given the card is red:", prob_heart_given_red, "\n")
```

Output :

```
> source("C:\\Users\\ADMIN\\Desktop\\conditional.R")
Probability of drawing a heart given the card is red: 0.5
```

Inference:

The probability of drawing a heart given that the card is red is 0.5. This demonstrates the application of conditional probability in determining outcomes based on given information

LAB Exercise 8 :

Practical based on Geometrical Probabili

Problem : Buffon's Needle is a famous problem in probability. You drop a needle of length l onto a floor with parallel lines spaced **d** apart. The goal is to estimate the probability that the needle will cross one of the lines.

For simplicity, let the needle length **l = d**, which is the simplest case. The probability of crossing a line is given by the formula **P= π/2.**

We'll simulate this experiment and estimate the probability by dropping a needle many times.

Aim: To estimate the probability that a needle will cross a line in Buffon's Needle problem using geometrical probability.

Procedure:

1. Set the number of experiments (num_trials) for simulation.
2. Assume the length of the needle l = d (distance between lines).
3. For each trial, randomly generate the distance from the needle's center to the nearest line and the angle of the needle.
4. Check if the needle crosses a line using the condition: distance <= (l/2) * sin(angle).
5. Estimate the probability as the ratio of successful crossings to the total number of trials.

R Code :

```
# Buffon's Needle Simulation
set.seed(123)
num_trials <- 10000
l <- 1 # Needle length
d <- 1 # Distance between lines
crossings <- 0

for (i in 1:num_trials) {
distance_to_nearest_line <- runif(1, 0, d / 2)
angle <- runif(1, 0, pi / 2)
if (distance_to_nearest_line <= (l / 2) * sin(angle)) {
```

```
crossings <- crossings + 1
  }
}
estimated_probability <- crossings / num_trials
cat("Estimated Probability from Simulation:", estimated_probability, "\n")
cat("Theoretical Probability (pi / 2):", pi / 2, "\n")
```

Output :

```
> source("C:\\Users\\ADMIN\\Desktop\\geometric.R")
Estimated Probability from Simulation: 0.6358
Theoretical Probability (pi / 2): 1.570796
```

Inference:

The Buffon's Needle simulation estimates the probability of crossing a line when dropping a needle onto a floor with parallel lines. The estimated probability should be close to the theoretical value of pi / 2, demonstrating the application of geometrical probability.

Course-105ST24: PRACTICAL-I Statistical Computing Using R -------Lab

LAB Exercise1:

Computing of Mean, Median, Geometric Mean and SD

Problem: Write a R code for computing Mean, Median, Geometric Mean and SD for the frequency distribution.

Execute R-code for the following data

Marks	15-19	20-24	25-29	30-34	35-39	40-44	45-49	50-54	55-59	60-64	65-69
No. of Students	9	11	10	44	45	54	37	36	18	5	1

Aim: To compute Mean, Median, Geometric Mean and Standard Deviation for a grouped frequency distribution using R programming.

Procedure:

From the given frequency distribution, we can compute mean, median, GM and SD by the following

i. $$Mean = \frac{\sum_{i=1}^{n} f_i x_i}{N}; \text{ where } N = \sum_{i=1}^{n} f_i$$

ii. The following are the step to find median

1. Arrange the data in ascending or descending order.
2. Find the cumulative frequency (cf) by adding the frequencies one by one.
3. Find $N/2$, where N is the total sum of the frequencies.
4. Find the cumulative frequency that is just greater than $N/2$.
5. The corresponding value of the variable is the median.

iii. $$Geometric\ Mean = \frac{\sum_{i=1}^{n} f_i \log x_i}{N}; \text{ where } N = \sum_{i=1}^{n} f_i$$

iv. $$\text{Standard Deviation } = \frac{\sum_{i=1}^{n} f_i x_i^2}{N} - (mean)^2$$

R Code:

```
# Data Input
marks_intervals<-c("15-19","20-24","25-29","30-34","35-39",
"40-44","45-49","50-54","55-59","60-64","65-69")
frequencies <-c(9,11,10,44,45,54,37,36,18,5,1)
```

```
# Calculate Midpoints
lower_limits<-seq(15,65, by =5)
upper_limits<-seq(19,69, by =5)
midpoints <-(lower_limits+upper_limits)/2

# Calculate Mean
mean_val<-sum(midpoints * frequencies)/sum(frequencies)

# Calculate Median
cumulative_freq<-cumsum(frequencies)
N <-sum(frequencies)
median_class_index<-which(cumulative_freq>= N /2)[1]
L <-lower_limits[median_class_index]
f <- frequencies[median_class_index]
cf_prev<-ifelse(median_class_index>1,cumulative_freq[median_class_index-
1],0)
w <-upper_limits[1]-lower_limits[1]
median_val<- L +((N /2-cf_prev)/ f)* w

# Calculate Geometric Mean
geometric_mean<-exp(sum(frequencies *log(midpoints))/sum(frequencies))

# Calculate Standard Deviation
mean_diff_squared<-(midpoints -mean_val)^2
variance <-sum(mean_diff_squared* frequencies)/sum(frequencies)
sd_val<-sqrt(variance)

# Output Results
cat("Mean:",mean_val,"\n")
cat("Median:",median_val,"\n")
cat("Geometric Mean:",geometric_mean,"\n")
cat("Standard Deviation:",sd_val,"\n")
```

Output:

```
Mean: 40.81481
Median:41.18519
Geometric Mean: 39.35265
Standard Deviation: 10.29576
```

LAB Exercise 2:

One Sample t-Test

Problem: A random sample of 10 students had the following I.Q.

70, 120, 110, 101, 88, 85, 95, 98, 107, 100.

Do this data support the assumptions of a population mean I.Q. of 100?

Aim: To test whether the sample is drawn from the population mean I.Q. of 100 or not using R programming.

Procedure:

Given sample size is less than 30.

Hence, we apply t-test.

Null Hypothesis (H_0) : The sample is drawn from the assumed population mean I.Q. of 100

i.e., H_0: $\mu = 100$

Alternative Hypothesis (H_1) : The sample is not drawn from the assumed population mean I.Q. of 100

i.e., H_1: $\mu \neq 100$

Choose level of significance $\alpha = 5\%$

Under null hypothesis H_0, the statistic is given by

$$t = \frac{\overline{x} - \mu}{s / \sqrt{n}} \sim t_{(n-1,\alpha)}$$

Where n= sample size, $\overline{x}$ = mean of the sample and $s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$

If $t_{tab} \leq t_{(n-1,\alpha)}$ then we accept the null hypothesis otherwise we reject the null hypothesis.

R Code :

```
# Test the significance of one sample t-test for mean with real-time data
input
cat("\nTest the significance of one sample t-test for mean\n");
# Define Null and Alternative Hypotheses
H0 <- readline("Define Null Hypothesis (e.g., 'The population mean is
100'): ")
H1 <- readline("Define Alternative Hypothesis (e.g., 'The population mean
is not 100'): ")

# Input real-time data
cat("\nEnter data values: ");
data <- scan()
```

```
# Compute statistics from data
n <- length(data)  # Sample size
xbar<- mean(data)  # Sample mean
sd<- sd(data)  # Sample standard deviation

# Input Population mean and significance level
cat("\nEnter Population mean (mu): ");
mu <- as.numeric(readline())
cat("\nEnter Level of significance (alpha): ");
alpha <- as.numeric(readline())

# Calculate the t-score
t_cal<- abs((xbar - mu) / (sd / sqrt(n)))

# Degrees of freedom
df <- (n - 1)

# Critical t-value (two-tailed)
t_tab<- abs(qt(1 - alpha / 2, df))

# Display results
cat("\nCalculated t-score =", t_cal)
cat("\nCritical t-value at alpha =", alpha, " is ", t_tab)

# Decision
if (t_cal<= t_tab) {
cat("\nWe accept the Null Hypothesis\n")
cat("\ni.e., ", H0, "\n")
} else {
cat("\nWe reject the Null Hypothesis\n")
cat("\ni.e., ", H1, "\n")
}
```

Input:

```
Test the significance of one sample t-test for mean
Define Null Hypothesis (e.g., 'The population mean is 100'): The population
mean is 100
Define Alternative Hypothesis (e.g., 'The population mean is not 100'): The
population mean is not 100
```

```
Enter data values: 1: 70
2: 120
3: 110
4: 101
5: 88
6: 85
7: 95
8: 98
9: 107
10: 100
11:
Read 10 items

Enter Population mean (mu): 100

Enter Level of significance (alpha): 0.05
```

Output:

```
Calculated t-score = 0.584569
Critical t-value at alpha = 0.05  is  2.262157
We accept the Null Hypothesis

i.e.,  The population mean is 100
```

LAB Exercise 3:

Two Sample t-Test

Problem: Below are given the gain in weights (in kgs) of pigs fat of two digits A and B

Diet A 25 32 30 34 24 14 32 24 30 31 35 25

Diet B 44 34 22 10 47 31 40 30 32 35 18 21 35 29 22

Aim: To test whether there is any significant difference between the gain weights of the two diets A and B.

Procedure:

Given two samples sizes less than 30.

Hence, we apply t-test.

Null Hypothesis (H_0) : There is no significant difference between the gain weights of two diets A and B.

Alternative Hypothesis (H_1) : There is a significant difference between the gain weights of two diets A and B.

Choose level of significance $\alpha = 5\%$

Under null hypothesis H_0, the statistic is given by

$$t = \frac{\bar{x} - \bar{y}}{s\Big/\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{(n_1+n_2-2,\alpha)}$$

Where $\bar{x}$ = mean of the sample diet A, $\bar{y}$ = mean of the sample diet B and $s^2 = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_1 - 2}$

where s_1^2 =sample variance of diet A, $s_1^2 = \frac{1}{n_1}\sum_{i=1}^{n}(x_i - \bar{x})^2$, n_1 = sample size of diet A

s_2^2 =sample variance of diet B, $s_2^2 = \frac{1}{n_2}\sum_{i=1}^{n}(y_i - \bar{y})^2$, n_2 = sample size of diet B.

If $t_{tab} \le t_{(n_1+n_2-2,\alpha)}$ then we accept the null hypothesis otherwise we reject the null hypothesis.

R Code:

```
cat("\n### Test the Significance of Two-Sample T-Test for Means ###\n")

# Define Null and Alternative Hypotheses
H0 <- readline("Define Null Hypothesis (e.g., 'The means of the two samples
are equal'): ")
```

```
H1 <- readline("Define Alternative Hypothesis (e.g., 'The means of the two
samples are not equal'): ")

# Input real-time data
cat("\nEnter data values for the first sample : ")
data1 <- scan()

cat("\nEnter data values for the second sample : ")
data2 <- scan()

# Compute statistics from data
n1 <- length(data1)
n2 <- length(data2)
xbar<- mean(data1)
ybar<- mean(data2)
sd1 <- sd(data1)
sd2 <- sd(data2)

cat("\nEnter the Level of significance (e.g., 0.05): ")
alpha <- as.numeric(readline())

#pooled sample variance
s=sqrt((n1*sd1^2+n2*sd2^2)/(n1+n2-2))
# Calculate the t-score
t_cal<- abs((xbar - ybar) / (s/sqrt((1 / n1) + (1 / n2))))

# Degrees of freedom
df <- n1+n2-2

# Critical t-value (two-tailed)
t_tab<- abs(qt(1 - alpha / 2, df))

# Display results
cat("\n### Results ###\n")
cat("First Sample Mean:", xbar, "\n")
cat("Second Sample Mean:", ybar, "\n")
cat("First Sample Standard Deviation:", sd1, "\n")
cat("Second Sample Standard Deviation:", sd2, "\n")
cat("Calculated t-score (t_cal):", t_cal, "\n")
cat("Degrees of Freedom (df):", round(df, 2), "\n")
cat("Critical t-value (t_tab) at alpha =", alpha, ":", t_tab, "\n")
```

```
# Decision
if (t_cal<= t_tab) {
cat("\nWe accept the Null Hypothesis\n")
cat("\ni.e.,", H0, "\n")
} else {
cat("\nWe reject the Null Hypothesis\n")
cat("\ni.e.,", H1, "\n")
}
cat("\n")
```

Input:

```
### Test the Significance of Two-Sample T-Test for Means ###
Define Null Hypothesis (e.g., 'The means of the two samples are equal'):
The means of the two samples are equal
Define Alternative Hypothesis (e.g., 'The means of the two samples are not
equal'): The means of the two samples are not equal

Enter data values for the first sample : 1: 25
2: 32
3: 30
4: 34
5: 24
6: 14
7: 32
8: 24
9: 30
10: 31
11: 35
12: 25
13:
Read 12 items
Enter data values for the second sample (separated by spaces): 1: 44
2: 34
3: 22
4: 10
5: 47
6: 31
7: 40
```

```
8: 30
9: 32
10: 35
11: 18
12: 21
13: 35
14: 29
15: 22
16:
Read 15 items

Enter the Level of significance (e.g., 0.05): 0.05
```

Output:

```
### Results ###
First Sample Mean: 28
Second Sample Mean: 30
First Sample Standard Deviation: 5.877538
Second Sample Standard Deviation: 10.03565
Calculated t-score (t_cal): 0.08826753
Degrees of Freedom (df): 25
Critical t-value (t_tab) at alpha = 0.05 : 2.059539

We accept the Null Hypothesis

i.e., The means of the two samples are equal
```

LAB Exercise 4:

Newton – Raphson Method

Problem: Solve the equation $x^3-2x-5=0$ using Newton – Raphson method by writing necessary R-code and executing the same.

Aim: To solve the equation $x^3-2x-5=0$ using Newton – Raphson method in R.

Procedure:

Newton-Raphson Method:

- This iterative method is based on the formula: $x_{n+1} = x_n - \frac{f(x)}{f'(x)}$ where:
  - f(x)= x^3-2x-5 is the function to be solved.
  - f′(x) = $3x^2-2$ is the derivative of f(x).
  - x_n is the current approximation, and x_{n+1} is the next approximation.

•**Steps**:

- Start with an initial guess x_0.
- Compute f(x) and f′(x) for the current value of x.
- Update x using the formula above.
- Repeat the steps until the difference between successive approximations is smaller than a pre-defined tolerance.

R Code :

```
# Demonstration program for NR method
# f(x)= x^3-2*x-5
f=function(x) return(x^3-2*x-5)
# df(x)= 3*x^2-2
df=function(x) return(3*x^2-2)

# Newton-Raphson method to solve f(x)=0
cat("\n\n Enter initial vaur of X:")
x0=scan()
for(n in 1:100)
{
xn=x0-f(x0)/df(x0)
if(abs(xn-x0)<0.000001) break();
x0=xn;
cat("\n \n Iteration ",n,":x=",x0);
```

```
}
if(n>=100) cat("\n NR method diverges") else cat("\n Final solution:
x=",xn);
cat("\n ## End of computation ## \n")
```

Input:

```
Enter initial vaur of X:1: 2.5
2:
Read 1 item
```

Output:

```
Iteration  1 :x= 2.164179

Iteration  2 :x= 2.097135

Iteration  3 :x= 2.094555

Iteration  4 :x= 2.094551
 Final solution: x= 2.094551
 ## End of computation ##
```

LAB Exercise 5:

Completely Randomized Design (CRD)

Problem: The weights in gm of a number of copper wires, each of length 1 meter, were obtained. These are classified according to the die from which they come.

Die No.				
I	II	III	IV	V
1.30	1.28	1.32	1.31	1.30
1.32	1.35	1.29	1.29	1.32
1.36	1.33	1.31	1.33	1.30
1.35	1.34	1.28	1.31	1.33
1.32	NA	1.33	1.32	NA
1.37	NA	1.30	NA	NA

Setup an analysis of variance table to test the significance of the difference between the weights due to different dies. Compute CRD analysis in R.

Aim: To analyse the given data by using CRD.

Procedure:

Null hypothesis H_0: The means of various treatments effects are homogeneous

i.e. H_0: $\mu_1 = \mu_2 = \mu_3 = \mu$

Alternative hypothesis H_1: The means of various treatments effects are not homogeneous

i.e. H_1: $\mu_1 \neq \mu_2 \neq \mu_3 \neq \mu$

Row sum of squares, RSS= $\sum_{i=1}^{k}\sum_{j=1}^{n_j} x_{ij}^2$

Grand Total, G= $\sum_{i=1}^{k}\sum_{j=1}^{n_j} x_{ij} = \sum_{i=1}^{k} T_i$

N= Total number of observations.

Correction factor (CF)= $\frac{G^2}{N}$

Total sum of squares, TSS=RSS-CF

Sum of squares due to treatments $SST = \sum_{i=1}^{k}\frac{T_i^2}{n_i} - CF$

Sum of squares due to error, SSE=TSS-SST

ANOVA table:

Source of Variation	Degrees of freedom	S.S.	M.S.S.	F-ratio
Treatments	k-1	SST	$MSST = \frac{SST}{k-1}$	$Fcal = \frac{MSST}{MSSE} \sim F_{(\alpha\%,(k-1),(N-k))}$
Error	N-K	SSE	$MSSE = \frac{SSE}{N-k}$	-
Total	N-1	TSS	-	

If Fcal is less than or equal to $F_{(\alpha\%,(k-1),(N-k))}$ then we accept the null hypothesis, otherwise we reject the null hypothesis.

R Code :

```
# Read data from a CSV file
data <- read.csv("crd.csv", header = TRUE)

# Display the given CRD data
cat("\nThe given CRD data is:\n")
print(data)

# Calculate and display treatment means
cat("\nTreatment means:\n")
print(colMeans(data, na.rm = TRUE)) # Use colMeans with na.rm = TRUE to
handle NA values

# Reshape the data into a long format for ANOVA
crd<- stack(data)
colnames(crd) <- c("yield", "Treatments")

# Perform ANOVA
crd.anova<- aov(yield ~ Treatments, data = crd)
cat("\nANOVA results:\n")
print(summary(crd.anova))
anova_summary<-summary(crd.anova)
# Extract degrees of freedom and F-value
df_treatments<- anova_summary[[1]][["Df"]][1]        # Degrees of freedom
for Treatments
df_residuals<- anova_summary[[1]][["Df"]][2]         # Degrees of freedom
```

```
for Residuals
f_calculated<- anova_summary[[1]][["F value"]][1]    # Calculated F-value

# Significance level
alpha <- 0.05

# Calculate critical F-value (table value)
f_critical<- qf(1 - alpha, df_treatments, df_residuals)

# Display calculated F-value and critical F-value
cat("\nComparison of Calculated F-value with Table Value:\n")
cat("Calculated F-value:", f_calculated, "\n")
cat("Critical F-value (alpha =", alpha, "):", f_critical, "\n")

# Hypothesis Testing Conclusion
if (f_calculated>f_critical) {
cat("\n Conclusion: Reject the null hypothesis (significant differences
between treatments).\n")
} else {
cat("\n Conclusion: Fail to reject the null hypothesis (no significant
differences between treatments).\n")
}
```

Output:

```
The given CRD data is:
     I   II  III   IV    V
1 1.30 1.28 1.32 1.31 1.30
2 1.32 1.35 1.29 1.29 1.32
3 1.36 1.33 1.31 1.33 1.30
4 1.35 1.34 1.28 1.31 1.33
5 1.32   NA 1.33 1.32   NA
6 1.37   NA 1.30   NA   NA

Treatment means:
       I       II      III       IV        V
1.336667 1.325000 1.305000 1.312000 1.312500

ANOVA results:
            Df   Sum Sq   Mean Sq F value Pr(>F)
Treatments   4 0.003598 0.0008994    1.81  0.167
Residuals   20 0.009938 0.0004969
5 observations deleted due to missingness
```

```
Comparison of Calculated F-value with Table Value:
Calculated F-value: 1.809995
Critical F-value (alpha = 0.05 ): 2.866081
```

Conclusion: Fail to reject the null hypothesis (no significant differences between treatments).

LAB Exercise 6:

Randomized Block Design (RBD)

Problem: Setup a table of analysis of variance for yields of three strains of wheat planted in five randomized blocks.

Strains	Blocks				
A	20	21	23	16	20
B	18	20	17	15	25
C	25	28	22	28	32

Write the necessary R code for RBD analysis

Aim: To analyse the data by using RBD.

Procedure:

Null hypothesis H_{0t}: The means of various treatments effects are homogeneous

i.e. H_{0t}: $\mu_1 = \mu_2 = \mu_3 = \mu$

Null hypothesis H_{0b}: The means of various block effects are homogeneous

i.e. H_{0b}: $\beta_1 = \beta_2 = \beta_3 = \beta$

Alternative hypothesis H_{1t}: The means of various treatments effects are not homogeneous

i.e. H_{1t}: $\mu_1 \neq \mu_2 \neq \mu_3 \neq \mu$

Alternative hypothesis H_{1b}: The means of various treatments effects are not homogeneous

i.e. H_{1b}: $\beta_1 \neq \beta_2 \neq \beta_3 \neq \beta$

Grand Total, G= $\sum_{i=1}^{k}\sum_{j=1}^{n_j} x_{ij} = \sum_{i=1}^{k} T_i$

N= Total number of observations.

Correction factor (CF)= $\frac{G^2}{N}$

Total sum of squares, TSS=RSS-CF

Row sum of squares, RSS= $\sum_{i=1}^{k}\sum_{j=1}^{n_j} x_{ij}^2$

Sum of squares due to treatments $SST = \sum_{i=1}^{k} \frac{T_i^2}{n_i} - CF$

Sum of squares due to Blocks $SSB = \sum_{j=1}^{nk} \frac{T_j^2}{n_j} - CF$

Sum of squares due to error, SSE=TSS-SST-SSB

ANOVA table:

Source of Variation	Degrees of freedom	S.S.	M.S.S.	F-ratio
Treatments	t-1	SST	$MSST = \frac{SST}{t-1}$	$Ftcal = \frac{MSST}{MSSE} \sim F_{(\alpha\%,(t-1),(t-1)(b-1))}$
Blocks	b-1	SSB	$MSSB = \frac{SSB}{B-1}$	$Fbcal = \frac{MSSB}{MSSE} \sim F_{(\alpha\%,(b-1),(t-1)(b-1))}$
Error	(t-1)(b-1)	SSE	$MSSE = \frac{SSE}{(t-1)(b-1)}$	-
Total	N-1	TSS	-	

If Ftcal is less than or equal to $F_{(\alpha\%,(t-1),(t-1)(b-1))}$ then we accept the null hypothesis H_{0t}, otherwise we reject the null hypothesis H_{0t}.

If Fbcal is less than or equal to $F_{(\alpha\%,(b-1),(t-1)(b-1))}$ then we accept the null hypothesis H_{0b}, otherwise we reject the null hypothesis H_{1b}.

R Code :

```
# Read data from a CSV file
data <- read.csv("rbd.csv", header = TRUE)

# Display the given RBD data
cat("\nThe given RBD data is:\n")
print(data)

# Set row names as block labels and remove the first column (block names)
rownames(data) <- data[,1]
data <- data[,-1] # Remove the first column (block names)

# Calculate and display treatment means
cat("\nTreatment means:\n")
treatment_means<- colMeans(data, na.rm = TRUE) # Use colMeans with na.rm =
TRUE
print(treatment_means)

# Calculate and display block means
cat("\nBlock means:\n")
block_means<- rowMeans(data, na.rm = TRUE) # Use rowMeans with na.rm = TRUE
```

```
print(block_means)

cat("\n\n")

# Reshape the data into a long format for ANOVA
rbd<- stack(data)
blocks <- rep(rownames(data), ncol(data)) # Repeat block names for each
treatment
rbd<- cbind(rbd, blocks) # Combine the reshaped data with block labels
names(rbd) <- c("yield", "Treatments", "Blocks") # Rename the columns

# Display the reshaped data
cat("\nThe reshaped RBD data is:\n")
print(rbd)

# Perform ANOVA for RBD
fit <- aov(yield ~ Treatments + Blocks, data = rbd)

# Display the ANOVA table
cat("\nANOVA table for RBD:\n")
print(summary(fit))

# Extract degrees of freedom and F-values from ANOVA table
anova_summary<- anova(fit)
treatment_df<- anova_summary$Df[1]
block_df<- anova_summary$Df[2]
residual_df<- anova_summary$Df[3]
treatment_F<- anova_summary$`F value`[1]
block_F<- anova_summary$`F value`[2]

# Critical F-values from the F-distribution
alpha <- 0.05
treatment_F_critical<- qf(1 - alpha, treatment_df, residual_df)
block_F_critical<- qf(1 - alpha, block_df, residual_df)

# Compare calculated F-values with table (critical) values
cat("\nComparison of F-values with Table Values:\n")
cat(sprintf("Treatments: Calculated F = %.3f, Critical F = %.3f\n",
treatment_F, treatment_F_critical))
cat(sprintf("Blocks: Calculated F = %.3f, Critical F = %.3f\n", block_F,
block_F_critical))
```

```
if (treatment_F>treatment_F_critical) {
cat("\nTreatments: Reject the null hypothesis (significant differences
among treatments).\n")
} else {
cat("\nTreatments: Fail to reject the null hypothesis (no significant
differences among treatments).\n")
}

if (block_F>block_F_critical) {
cat("\nBlocks: Reject the null hypothesis (significant differences among
blocks).\n")
} else {
cat("\nBlocks: Fail to reject the null hypothesis (no significant
differences among blocks).\n")
}
```

Output:

```
The given RBD data is:
X  I II III IV  V
1 A 20 21  23 16 20
2 B 18 20  17 15 25
3 C 25 28  22 28 32

Treatment means:
       I       II      III       IV        V
21.00000 23.00000 20.66667 19.66667 25.66667

Block means:
A  B  C
20 19 27

The reshaped RBD data is:
   yield Treatments Blocks
1     20          I      A
2     18          I      B
3     25          I      C
4     21         II      A
5     20         II      B
6     28         II      C
7     23        III      A
8     17        III      B
```

```
9     22        III      C
10    16         IV      A
11    15         IV      B
12    28         IV      C
13    20          V      A
14    25          V      B
15    32          V      C

ANOVA table for RBD:
            Df Sum Sq Mean Sq F value Pr(>F)
Treatments   4     68      17   1.889 0.2059
Blocks       2    190      95  10.556 0.0057 **
Residuals    8     72       9
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Comparison of F-values with Table Values:
Treatments: Calculated F = 1.889, Critical F = 3.838
Blocks: Calculated F = 10.556, Critical F = 4.459

Treatments: Fail to reject the null hypothesis (no significant differences
among treatments).

Blocks: Reject the null hypothesis (significant differences among blocks).
```

LAB Exercise 7:

Fitting Regression Lines

Problem: A group of students recorded their study hours and corresponding exam scores:

Study Hours: 2, 3, 5, 7, 8, 10, 12, 14

Exam Scores: 50, 55, 65, 70, 75, 80, 90, 95

Analyze the relationship between study hours and exam scores by fitting a regression line.

Aim: To fit a simple linear regression line to the data and interpret the results.

Procedure:

1. Input the given study hours and exam scores into R.
2. Use the `lm()` function to fit a regression model.
3. Summarize the regression results to find the regression equation.
4. Visualize the data with a scatter plot and add the regression line.

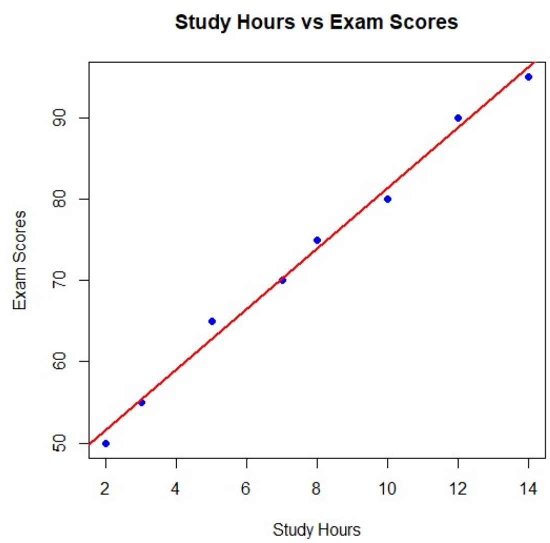
R Code:

```
# Given Data
study_hours<-c(2,3,5,7,8,10,12,14)
exam_scores<-c(50,55,65,70,75,80,90,95)

# Fit Regression Model
model <-lm(exam_scores~study_hours)

# Summary of Model
summary(model)

# Plotting
plot(study_hours,exam_scores,pch=16, col="blue",
     main="Study Hours vs Exam Scores",xlab="Study Hours",ylab="Exam
Scores")
abline(model, col="red",lwd=2)
```

Output

```
Call:
lm(formula = exam_scores ~ study_hours)
Residuals:
    Min      1Q  Median      3Q     Max
-1.6087 -1.2128 -0.2507  1.1433  2.2493
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  44.1807     1.1578   38.16 2.16e-08 ***
study_hours   3.7140     0.1347   27.57 1.51e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.511 on 6 degrees of freedom
Multiple R-squared:  0.9922,    Adjusted R-squared:  0.9909
F-statistic: 760.2 on 1 and 6 DF,  p-value: 1.505e-07
```



LAB Exercise 8:

High-Level Plotting - Creating Different Types of Plots

Problem: Consider a dataset containing the monthly sales (in thousands) of a product in different regions:

Regions: North, South, East, West

Sales: 50, 60, 45, 70

Create the following visualizations to analyze the data:

1. Histogram of sales.
2. Bar plot of regional sales.
3. Box plot to compare the distribution of sales.

Aim: To create various types of plots for visualizing sales data using high-level plotting functions.

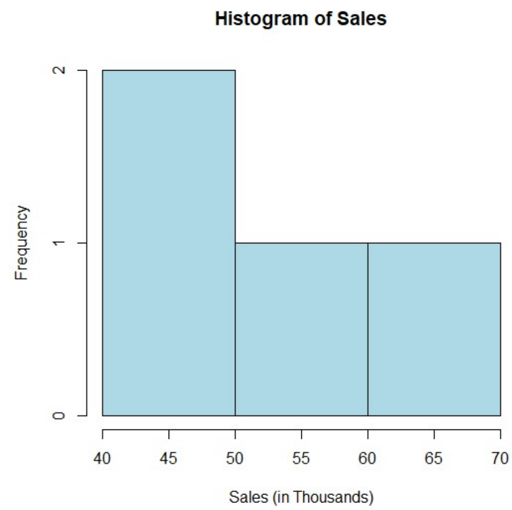
Procedure:

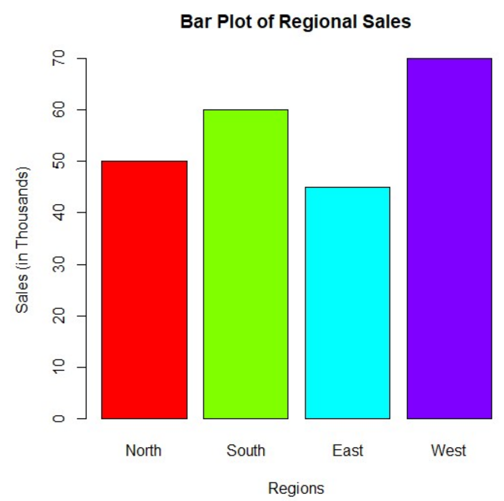
1. Use the provided sales data.
2. Create a histogram to view the sales distribution.
3. Use a bar plot to compare sales across regions.
4. Generate a box plot to summarize the sales data.

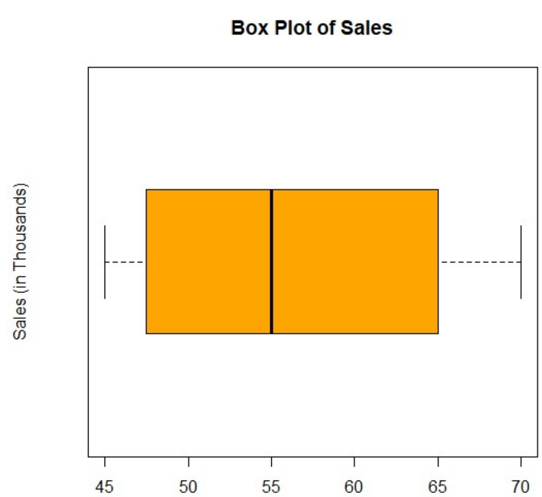
R Code

```
# Given Data
regions <-c("North","South","East","West")
sales <-c(50,60,45,70)

# Histogram
hist(sales, main="Histogram of Sales",xlab="Sales (in Thousands)",
col="lightblue", border="black")

# Bar Plot
barplot(sales,names.arg=regions, col=rainbow(4),
        main="Bar Plot of Regional Sales",xlab="Regions",ylab="Sales (in
Thousands)")

# Box Plot
boxplot(sales, main="Box Plot of Sales", col="orange",
ylab="Sales (in Thousands)", horizontal=TRUE)
```

Output:


Histogram of Sales
Frequency
0
1
2
40
45
50
55
60
65
70
Sales (in Thousands)


Bar Plot of Regional Sales
Sales (in Thousands)
0
10
20
30
40
50
60
70
North
South
East
West
Regions


Box Plot of Sales
Sales (in Thousands)
45
50
55
60
65
70

LAB Exercise 9:

Low-Level Plotting - Adding Elements to a Plot.

Problem: For the following dataset:

X values: 1, 2, 3, 4, 5, 6, 7, 8

Y values: 2, 4, 6, 8, 10, 12, 14, 16

Create a scatter plot and customize it by adding:

- A horizontal reference line at Y=10.
- A vertical reference line at X=5.
- Gridlines and custom points.

Aim: To demonstrate the use of low-level plotting functions for enhancing a plot.

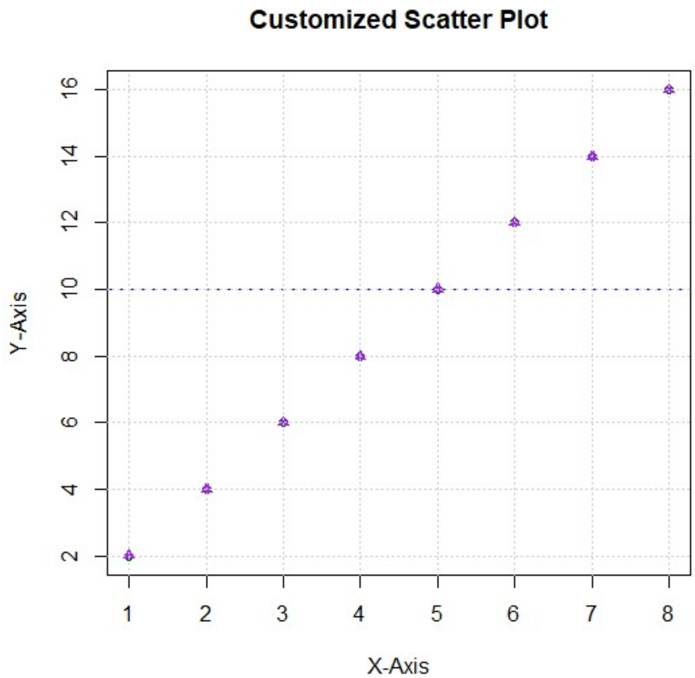
Procedure

1. Plot the given data using `plot()`.
2. Add elements using functions like `abline()` and `grid()`.

R Code

```
# Given Data
x <-c(1,2,3,4,5,6,7,8)
y <-c(2,4,6,8,10,12,14,16)

# Base Scatter Plot
plot(x, y,pch=19, col="darkgreen", main="Customized Scatter Plot",
xlab="X-Axis",ylab="Y-Axis")

# Adding Elements
abline(h=10, col="blue",lty=2)# Horizontal line at Y = 10
abline(v=5, col="red",lty=3)# Vertical line at X = 5
points(x, y, col="purple",pch=17)# Adding points
grid()# Adding gridlines
```

Output


Customized Scatter Plot
Y-Axis
X-Axis
2
4
6
8
10
12
14
16
1
2
3
4
5
6
7
8

LAB Exercise 10:

Q-Q Plot and Pie Chart

Problem: Given the following datasets:

1. Normal data: X={2,4,6,8,10,12,14,16}.
2. Categories: A (25%), B (35%), C (20%), D (20%).

Create:

1. A Q-Q plot to check normality for X.
2. A pie chart to visualize the categorical proportions.

Aim: To use Q-Q plots for distribution analysis and pie charts for proportion visualization.

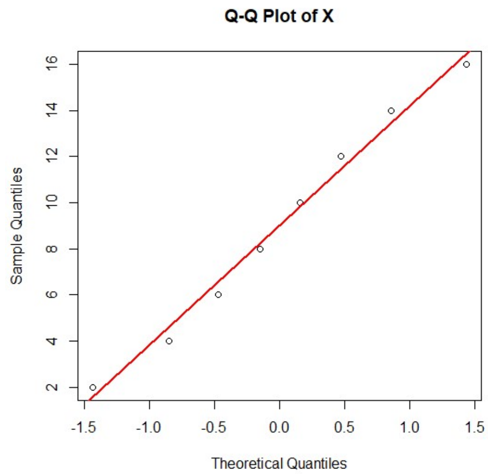
Procedure:

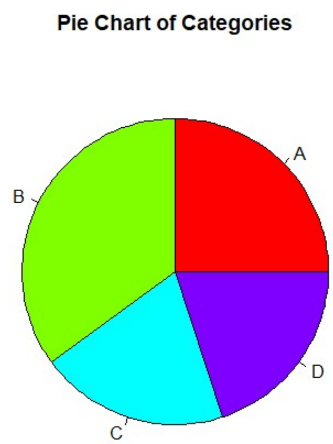
1. Generate a Q-Q plot for the provided numerical data.
2. Create a pie chart for the given categorical data.

R Code:

```
# Q-Q Plot
x <-c(2,4,6,8,10,12,14,16)
qqnorm(x, main="Q-Q Plot of X")
qqline(x, col="red",lwd=2)

# Pie Chart
categories <-c("A","B","C","D")
values <-c(25,35,20,20)
pie(values, labels=categories, main="Pie Chart of Categories",
col=rainbow(4))
```

Output:


Q-Q Plot of X
Sample Quantiles
Theoretical Quantiles
2
4
6
8
10
12
14
16
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5


Pie Chart of Categories
A
B
C
D