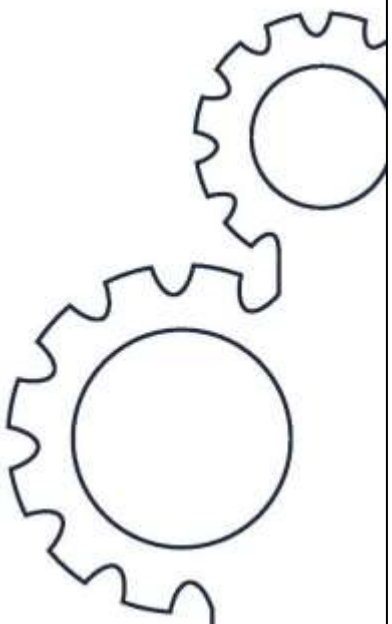
Kumaraguru
College of Technology
Coimbatore - 641049

U18MAI4201-
Statistical Lab using
R-Programming



Name:

Roll Number:

Department:

KUMARAGURU
college of technology
character is life

Certificate

Certified that this is the bonafide record work done by

Mr / Ms_____, (Reg No:_____) of

___ Year,__Semester of the Department of _____

during the Academic Year 2023–2024 (Even Semester) in the

_____ laboratory.

Faculty In charge

Submitted for the End Semester Practical Examination

Conducted on _____

Examiner I **Examiner II**

TABLE OF CONTENTS

Experiment number : 1 **Date:**

INTRODUCTION TO R PROGRAMMING

STEP 1: INTRODUCTION

OBJECTIVES OF THE EXPERIMENT

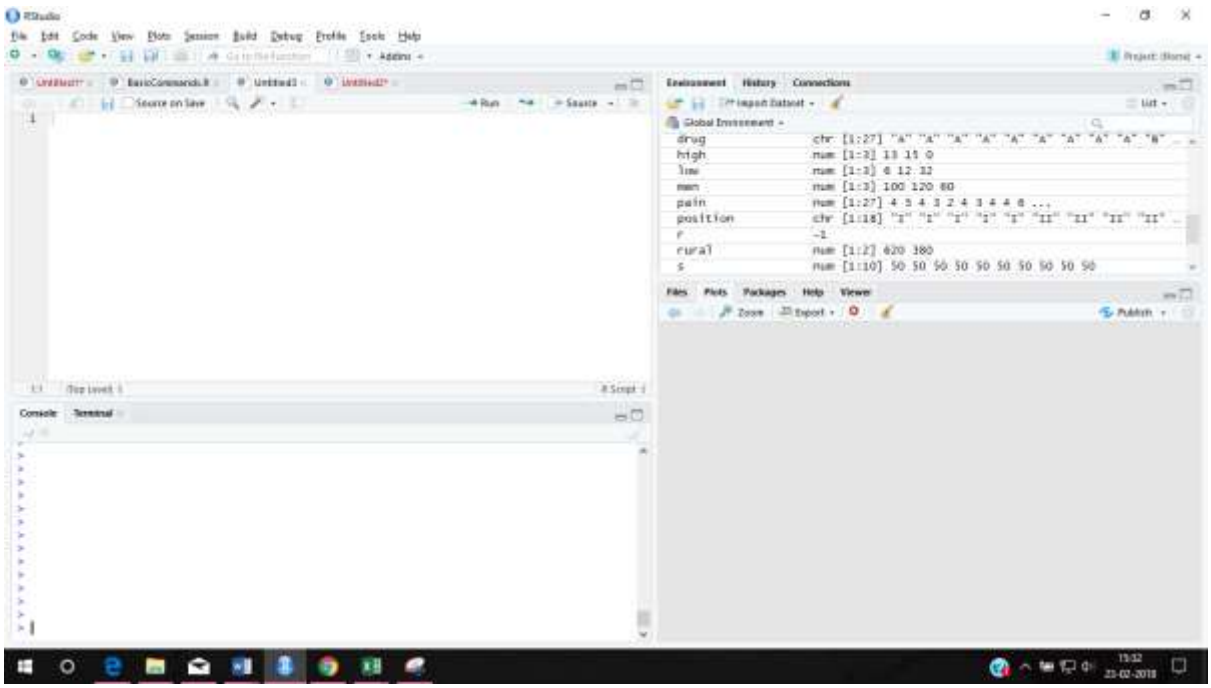
1. To understand the basics of R-Programming and R Studio
2. To understand the representations of basic data
3. To create a data frame using given data
4. To import data from a given MS-Excel file

STEP 2: ACQUISITION

I. INTRODUCTION TO R PROGRAMMING

- R is a programming language and software environment for statistical analysis, graphics representation and reporting. R was created by Ross Ihaka and Robert Gentleman at the University of Auckland, New Zealand
- This programming language was named R, based on the first letter of first name of the two R authors (Robert Gentleman and Ross Ihaka)
- R allows integration with the procedures written in the C, C++, .Net, Python or FORTRAN languages for efficiency.
- R is a well-developed, simple and effective programming language which includes conditionals, loops, user defined recursive functions and input and output facilities.
- R has an effective data handling and storage facility
- R provides a suite of operators for calculations on arrays, lists, vectors and matrices.
- R provides a large, coherent and integrated collection of tools for data analysis.
- R provides graphical facilities for data analysis and display either directly at the computer or printing at the papers.
- R is the world's most widely used statistics programming language.

The four windows



Editor	Command History (Environment)
Console	Instructions Packages Plot

Important note: R is case sensitive

R-objects:

There are many types of R-objects.

- Vectors
- Lists
- Matrices
- Arrays
- Factors
- Data Frames

Vector: a <- c(1,2,3,4,5,6)

(or) a = c(1,2,3,4,5,6)

Some basic commands

1. **To generate a sequence with common difference 1**

R code: seq(1,10)

Output : 1 2 3 4 5 6 7 8 9 10

2. **To generate a sequence with common difference 2**

R code :seq(1,15,by=2) :

Output: 1 3 5 7 9 11 13 15

3. **To find the square root of a number**

R code:

#square root y=2

sqrt(2) (or) x=sqrt(y)

Output:

[1] 1.414214 [1] 1.414214

4. **To perform addition of two numbers**

R-code		**R-codeR-code**	
a=2		c=a+b	c=2+3
b=3		a=2	c
c=a+b	(or)	b=3 (or)	
c		c	
Output		**OutputOutput**	
[1] 5		[1] 5	[1] 5

Data frames :

- Tabular data objects.
- Each column can contain different modes of data. The first column can be numeric while the second column can be character and third column can be logical.
- It is a list of vectors of equal length.
- Data Frames are created using the data.frame() function.

II. To create a data frame using given data

Procedure for doing the Experiment:

1.	Represent the various columns of the data frame by the vectors x, y, z ……
2.	A=data.frame(x,y,z,…..) creates the data frame.

Example

R code:

```
 A =data.frame(  name=c("A","B","C"),

gender = c("Male", "Male","Female"),

height = c(152, 171.5, 165),

weight = c(81,93, 78),

age =c(42,38,26))

 A

(OR)

name=c("A","B","C")

gender = c("Male", "Male","Female")

height = c(152, 171.5, 165)

weight = c(81,93, 78)

age =c(42,38,26)

 A=data.frame(name,gender,height,weight,age)

 A
```

Output

```
name gender height  weight age
1    A   Male    152.0     81   42
2    B   Male    171.5     93   38
3    C  Female  165.0     78   26
```

Task 1

Create a data frame from the following details regarding babies' frocks (Given: size, season. material, decoration, pattern type, price)

1. **L, spring, silk, embroidery, dot, 650**
2. **M, summer, chiffon, bow, print, 275**
3. **M, summer, cotton, null, animal, 380**
4. **M, Winter, cotton, null, patchwork, 450**
5. **L, autumn, linen, ruffles, animal, 420**

R Code:

```
size =c("L","M","M","M","L")
        season=c("Spring","Summer","Summer","Winter","Autumn")
        material=c("silk","chiffon","cotton","cotton","linen")
decoration=c("Embroidery","Bow","Null","Null","Ruffles")
patterntype=c("dot","print","animal","patchwork","animal")
price=c(650,275,380,450,420)
        D=data.frame(size,season,material,decoration,patterntype,price)
        D
```

Output:

	size	season	material	decoration	pattern type	price
1	L	Spring	silk	Embroidery	dot	650
2	M	Summer	chiffon	Bow	print	275
3	M	Summer	cotton	Null	animal	380
4	M	Winter	cotton	Null	patchwork	450
5	L	Autumn	linen	Ruffles	animal	420

III. To import data from a given MS-Excel file

1.	**To locate current working directory**
	# Get and print current working directory.
	print(getwd( ))

2.	**To import data from Excel sheet** To import data from Excel sheet 'abc', first save the file as .csv(comma delimited) in the current working directory. Then execute the following command data = read.csv("abc.csv") data
3.	**$ symbol is used to extract a specific field.**

Example:
To import data from Excel sheet

To import data from Excel sheet ‘Testmarks’, first save the file as .csv (comma delimited) in the current working directory. Then execute the following command

```
data = read.csv("Testmarks.csv")

data
```

Output:

Output is

```
Sl.No Name IT.1 IT.II
1       1    A   26    32
2       2    B   25    25
3       3    C   19    31
4       4    D   14    26
5       5    E   25    28
6       6    F   32    32
7       7    G   29    42
8       8    H   25    26
9       9    I   31    38
10     10    J   35    39
11     11    K   33    31
12     12    L   35    36
```

To get the list of students who have passed in Internal test 1

```
pass= subset(data, IT.1 >= 25 )

print(pass)
```

Output is

```
Sl.No Name IT.1 IT.II
1       1    A   26    32
2       2    B   25    25
5       5    E   25    28
6       6    F   32    32
7       7    G   29    42
8       8    H   25    26
9       9    I   31    38
10     10    J   35    39
11     11    K   33    31
12     12    L   35    36
```

To get the list of students who have secured 30 or more marks in both tests

```
Good= subset(data, IT.1 >= 30&IT.II>=30 )

Good
```

Output

```
Sl.No Name IT.1 IT.II

6       6    F   32    32
9       9    I   31    38
10     10    J   35    39
11     11    K   33    31
12     12    L   35    36
```

STEP 3: PRACTICE/TESTING

1. **What is a data frame?**

2. **Mention some characteristics of a data frame.**

FACULTY ASSESSMENT

Description	**Max Marks Awarded**
Preparation	10
Conduct of Experiment & Result	10
Viva	10
Total	30
Faculty Signature	

Experiment number : 2 **Date:**

APPLICATION OF DESCRIPTIVE STATISTICS – MEAN, MEDIAN, MODE AND STANDARD DEVIATION

STEP 1: INTRODUCTION

OBJECTIVES OF THE EXPERIMENT

To find arithmetic mean,median, mode and standard deviation.

STEP 2: ACQUISITION

1. To find the Arithmetic Mean

```
A=c(54,55,53,56,52,52,58,49,50,51)
Mean1=mean(A)
Mean1
[1] 53
```

2. To find the Median

```
A=c(54,55,53,56,52,52,58,49,50,51)
Med=median(A)
Med
[1] 52.5
```

3. To find the mode

Create the function.

```
mode=function(x){
ux= unique(x)
ux[which.max(tabulate(match(x,ux)))]
}
# Find the mode of the numbers 2,1,2,3,1,2,3,4,1,5,5,3,2,3
x = c(2,1,2,3,1,2,3,4,1,5,5,3,2,3)
# Calculate the mode using the user function.
result= mode(x)
print(result)
```

3. To find the standard deviation

```
A=c(54,55,53,56,52,52,58,49,50,51)
Std=sd(A)
Std
Output:
[1] 2.788867
```

Task 1: To find the average set length in a sizing unit

The following set lengths are used in a sizing unit in a factory during a month. Compute the arithmetic mean and median: 1780, 1760, 1690, 1750, 1840, 1920, 1100, 1810, 1050, 1950.

R Code:

Output:

R Code:

Output:

Task 2: Find the average export of steel in a month from the data given below (in millions of kgs) using mean and median:

Jan'16	105.26
Feb'16	101.05
Mar '16	113.60
Apr'16	105.97
May'16	95.05
Jun'16	93.58
Jul'16	76.21
Aug'16	67.42
Sep'16	77.88
Oct'16	77.97
Nov'16	104.44
Dec'16	174.11

R-Code:

Output:

R-Code:

Output:

Task 3: To find the average export of raw cotton per year

The following list gives the export quantity of raw cotton (in million kg.) for five consecutive years 2012-2013 to 2016-17: 1945.63, 1864.69, 1093.11, 1297.27, 918.15. Find the mean and median.

To find the Arithmetic mean, median, standard deviation for a frequency distribution

Example

```
d=read.table(header=TRUE,text="Marks        Frequency
+                5          15
+                15         20
+                25         30
+                35         20
+                45         17
+                55         6")
```

```
d2= rep(d$Marks, d$Frequency)
multi.fun = function(x) {
c(mean = mean(x), median = median(x), sd = sd(x))
}
multi.fun(d2)
Output:
mean      median    sd
27.03704  25.00000   14.25792
```

Task 4

Find the mean and standard deviation of the frequency distribution:

x:	1	2	3	4	5	6	7
f:	5	9	12	17	14	10	6

Task 5

The following data related to the distance traveled by 520 villagers to buy their weekly requirements.

Miles Traveled: 2 4 6 8 10 13 14 16 18 20

No of Villagers: 38 104140 78 48 42 28 24 16 2

Calculate the arithmetic mean and median.

Task 6

Calculate the mean and standard deviation for the following:

Size	**: 6**	**7**	**8**	**9**	**10**	**11**	**12**
Frequency:	**3**	**6**	**9**	**13**	**8**	**5**	**4**

Task 7

Find the mean, median and mode for the following data.

14.8, 14.2, 13.8, 13.5, 14.0, 14.2, 14.3, 14.6, 13.9, 14.0, 14.1, 13.2, 13.0, 14.2, 13.5, 13.0, 12.8, 13.9, 14.8, 15.0, 12.8, 13.4, 13.2, 14.0, 13.8, 13.9, 14.0, 14.0, 13.9, 14.8

To import data from a given MS-Excel file and to find arithmetic mean, median, mode and standard deviation

1.	**To locate current working directory** # Get and print current working directory. print(getwd( ))
2.	**To import data from Excel sheet** To import data from Excel sheet ‘abc’, first save the file as .csv (comma delimited) in the current working directory. Then execute the following command data = read.csv("abc.csv")

	data
3.	**$ symbol is used to extract a specific field.**
4.	**mean(data$-----)**
5.	**Median(data$----)**

6.	**Mode** :
	mode = function(x) {
	ux = unique(x)
	ux[which.max(tabulate(match(x, ux)))]
	}
	x = data$ ------
	# Calculate the mode using the user function.
	v = mode(x)
	print(v)

7.	**Standard Deviation** z=sd(data$ ----) z
8.	summary(data)

STEP 3: PRACTICE/TESTING

1. **Define arithmetic mean. Write the formula for finding mean of discrete data and frequency distribution.**

2. **Define median.**

3. **Define mode.**

4. **Define Standard Deviation. Write its formula for discrete data and for frequency distribution**

FACULTY ASSESSMENT

Description	**Max Marks Awarded**
Preparation	10
Conduct of Experiment & Result	10
Viva	10
Total	30
Faculty Signature	

Experiment number : 3 **Date:**

APPLICATIONS OF CORRELATION AND REGRESSION

STEP 1: INTRODUCTION

OBJECTIVES OF THE EXPERIMENT

1. To construct the scatter plot and to visualize the relationship between two quantitative variables.
2. To find the correlation between two variables in a data set.
3. To find the coefficient of rank correlation between two variables in a data set by Spearman's method.
4. To determine the equations of the regression lines for variables and to predict the value of one variable when the value of the other variable is given.
5. To construct the regression plot for the given variables.

STEP 2: ACQUISITION

Procedure for doing the Experiment:

1.	**To construct the scatter plot with the variables x and y** x=c(a,b,....) y=c(l,m,....) plot(x,y, xlab = "....",ylab="…",xlim=c(0,10),ylim=c(0,25),col=c("…"),main="…..")
2.	**To find the correlation between x and y** x=c(a,b,....) y=c(l,m,....) r=cor(x,y) r
3.	**To find the Spearman's rank correlation coefficient between x and y** x=c(a,b,....) y=c(l,m,....) r=cor(x,y,method="spearman") r
4.	**To find regression line of y on x** regyx=lm(y~x) #lm stands for linear model regyx

5.	**To find regression line of x on y** regxy=lm(x~y) regxy **To construct the regression plot of y on x**
6.	plot(x,y) abline(lm(y ~ x),col="---")

Note:

- **i)** plot(y~x) --- creates a scatterplot of y versus x
- **ii)** regmodel = lm(y~x) --- fit a regression model
- **iii)** abline(lm(y~x)) --- adds regression line to plot

Example

Construct the scatter plot and also find the coefficient of correlation and Spearman's correlation coefficient between the ends per inch(X) and picks per inch (Y). Also find the two regression lines. Estimate the value of y when x = 26.

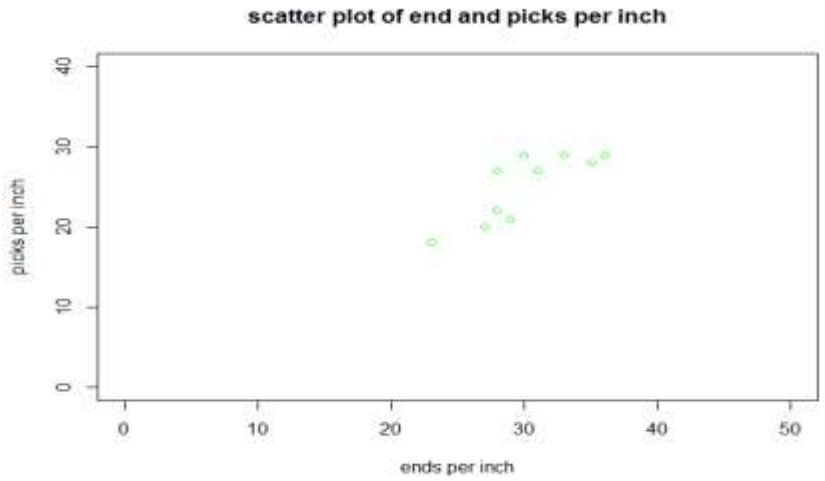
x	**23**	**27**	**28**	**28**	**29**	**30**	**31**	**33**	**35**	**36**
y	**18**	**20**	**22**	**27**	**21**	**29**	**27**	**29**	**28**	**29**

Solution:

R code:

```
x=c(23,27,28,28,29,30,31,33,35,36)
y=c(18,20,22,27,21,29,27,29,28,29)
plot(x,y,xlab ="ends per inch",ylab ="picks per
inch",xlim=c(0,50),ylim=c(0,40),col=c("green"),main="scatter plot of end and picks per
inch")
r=cor(x,y)
r
rank=cor(x,y,method="spearman")
rank
```

Scatter Plot:



Output:

Correlation Coefficient = 0.8176052
Spearman correlation coefficient= 0.9955947

Conclusion:
There is strong positive correlation between **ends per inch(X) and picks per inch(Y).**

To find the regression line of y on x
regyx=lm(y~x)

regyx

Output

Call:
lm(formula = y ~ x)

Coefficients:
(Intercept) x
-1.7391 0.8913
ie, regression line of y on x is y= -1.7391+0.8913x

To find the regression line of x on y:

regxy=lm(x~y)

regxy

Output:

Call:
lm(formula = x ~ y)

Coefficients:
(Intercept) y
11.25 0.75
ie, regression line of x on y is 11.25+0.75y

To find y when x=26

y1= - 1.7391+0.8913*26

y1

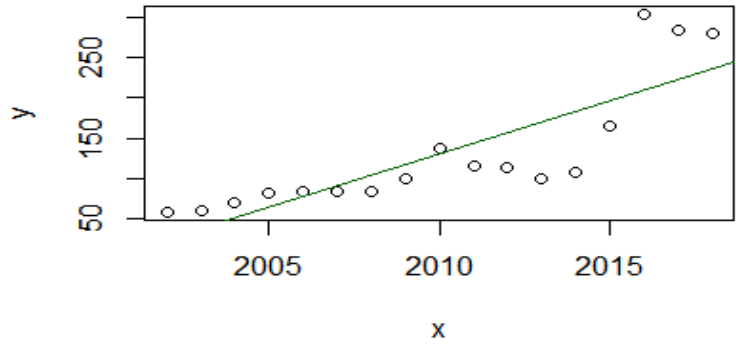
[1] 21.4347

Regression plot of y on x

R Code:

```
plot(x,y)
abline(lm(y ~ x),col="dark green")
```

Plot:



Task 1

Calculate the coefficient of correlation from the following figures relating to the consumption of fertilizer and the output of food grains in a district X:
Chemical fertilizer used (in metric tonnes):100,110,120,130,140,150,160,170,180,190,200,210,220,230
Output of food(in metric tonnes): 1000,1050,1080,1150,1200,1220,1300,1360,1420,1500,1600,1650,1650,1650
Also draw the scatter plot diagram for the above data and justify the result.
R-Code:

Output:

Scatter Plot

Task 2

Below are given the simple index numbers for the price of USB sound card for a number of years. Determine the scatter plot and correlation coefficient for the trend.

Year:2002,2003,2004,2005,2006,2007,2008,2009,2010,2011,2012,2013,2014,2015,2016,2017,2018

I.N:59,59.6,70,82.5,83.4,83.4,83.4,100,138.4,115.6,114.3,99.7,108.3,165,303.7,285.1,280.8

R-CODE:

Output:

Scatter Plot:

Task 3

Fifteen dishes in a cooking competition are ranked by 3 judges A, B, C in the following order.

A: 14,15,1,6,5,3,10,2,4,9,7,8,12,13,11

B:15,13,11,3,5,8,4,7,10,2,1,6,9,12,14

C:12,11,6,4,9,8,1,2,3,10,5,7,15,14,13

Find which pair of judges have the nearest approach to common taste in food.

R-CODE:

Output:

Conclusion:

Task 4:

The following data are related to the percentage of humidity and the warp breakage rate recorded for a week in a loom shed.

Percentage humidity	**54**	**85**	**86**	**50**	**42**	**75**	**65**	**56**
Warp breakage rate	**2.45**	**1.21**	**1.20**	**2.84**	**3.25**	**1.86**	**1.90**	**2.32**

Find two equations of lines of regression. In addition, find warp breakage rate if humidity percentage on a specific day is 60 and find percentage humidity required for the target warp breakage rate of 1.50%.

R-CODE:

```
#when x=60
y=4.91906-0.04351*60
y
#when y=1.5
y=111.03-22.03*1.50
y
```

OUTPUT:

Task 5:
From the following data, obtain the two regression equations:
Sales: 91,97,108,121,67,124,51,73,111, 57
Purchases: 71,75,69,97,70,91,39,61,80,47
Also compute the most likely purchase when sales = 150 and construct the regression plot of purchases on sales.

R-CODE:

```
#when x=90
y=14.8113+0.6132*90
y
plot()
abline(lm(pur~sal),col="red")
```

OUTPUT:

Plot:

Task 6:
Compute the two equations of the regression lines for the following data:
A panel of judges A and B graded seven debaters and independently awarded the following marks:

Marks by A:	**40**	**34**	**28**	**30**	**44**	**38**	**31**
Marks by B:	**32**	**39**	**26**	**30**	**38**	**34**	**28**

An eighth debater was awarded 36, marks by Judge A while Judge B was not present. If Judge B was also present, how many marks would you expect him to award to eighth debater assuming same degree of relationship exists in judgment?

R-CODE:

OUTPUT:

Task 7:
The following table gives the ages and blood pressure of 10 men.

Age (X):	**56**	**42**	**36**	**47**	**49**	**42**	**60**	**72**	**63**	**55**
Blood Pressure(Y):	**147**	**125**	**118**	**128**	**145**	**140**	**155**	**160**	**149**	**150**

Find (i) The two regression line equations.
(ii) Estimate the blood pressure of men whose age is 45 years
(iii) Estimate the age of men whose blood pressure is 172.
(iv) Construct the regression plot of blood pressure on age.

R-CODE:

OUTPUT:

Plot:

STEP 3: PRACTICE/TESTING

1. **Define correlation.**

2. **What are the various methods of studying correlation?**

3. **Explain scatter diagram.**

4. **Define regression.**

5. **What are regression lines? Write their equations.**

6. **Mention some properties of regression lines.**

FACULTY ASSESSMENT

Description	**Max Marks Awarded**
Preparation	10
Conduct of Experiment & Result	10
Viva	10
Total	30
Faculty Signature	

Experiment number : 4 **Date:**

APPLICATIONS OF NORMAL DISTRIBUTION

STEP 1: INTRODUCTION

OBJECTIVES OF THE EXPERIMENT

To predict values and compute probabilities using normal distribution

STEP 2: ACQUISITION

The normal distribution is defined by the following probability density function, where μ is the population mean and σ^2 is the variance.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

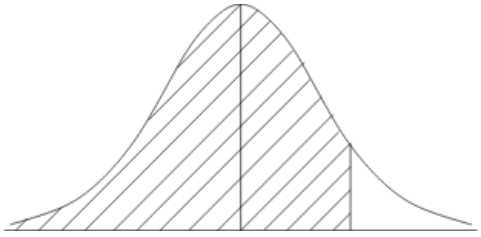
If a random variable X follows the normal distribution, then we write: $X \sim N(\mu, \sigma^2)$
The normal distribution with μ = 0 and σ = 1 is called the standard normal distribution, and is denoted as N(0,1).

Consider a normal distribution with mean μ and standard deviation σ

R-code for doing the Experiment:

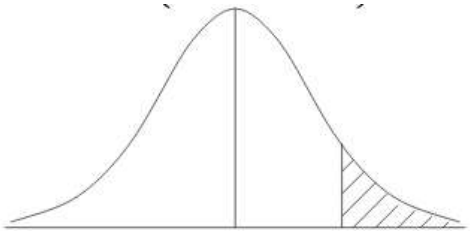
1.	To find $P(X < a) = P(-\infty < X < a)$ **R-code :** pnorm(a, mean = μ, sd = σ)
2.	To find $P(X > a) = P(a < X < \infty)$ R-code: pnorm(a, mean = μ, sd = σ, lower.tail = FALSE)
3.	To find $P(a < X < b)$ R-code: pnorm(b, mean = μ, sd = σ) - pnorm(a, mean = μ, sd = σ)

To find $P(X < a) = P(-\infty < X < a)$



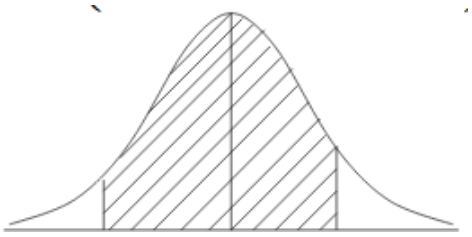
pnorm (a, mean = μ, sd = σ)

To find $P(X > a) = P(a < X < \infty)$



pnorm(a, mean = μ, sd = σ, lower.tail = FALSE)

To find $P(a < X < b)$



pnorm(b, mean = μ, sd = σ) - pnorm(a, mean = μ, sd = σ)

Note:
Use lower.tail=TRUE if you are finding the probability at the lower tail of a confidence interval or if you want to estimate the probability of values no larger than z.

Use lower.tail=FALSE if you aretrying to calculate probability at the upper confidence limit, or you want the probability of values z or larger.

Example
A certain type of storage battery lasts on the average 3.0 years with standard deviation of 0.5 year. Assuming that the battery lives are normally distributed, find the probability that a given battery will last

(i) less than 2.3 years (ii) more than 3.1 years (iii) between 2.5 and 3.5 years

Ans:

(i) pnorm(2.3, mean=3.0, sd=0.5)
[1] 0.08075666

(ii) pnorm(3.1, mean=3.0, sd=0.5, lower.tail=FALSE)
[1] 0.1586553

(iii) pnorm(3.5, mean=3.0, sd=0.5) - pnorm(2.5, mean=3.0, sd=0.5)
[1] 0.6826895

Task 1
Suppose the heights of men of a certain country are normally distributed with average 68 inches and standard deviation 2.5, find the percentage of men who are

(i) between 66 inches and 71 inches in height
(ii) approximately 6 feet tall (ie, between 71.5 inches and 72.5 inches)

Task 2

The mean yield for one acre plots is 662 kgs with S.D 32. Assuming normal distribution, how many one acre plots in a batch of 1000 plots. Would you expect to yield .

(i) Over 700 kgs

(ii) Below 650 kgs.

(Note: Find the respective probabilities and multiply the probabilities by the number of plots (= 1000) to get the final answers)

Task 3

A bore in picking element of a projectile loom part produced is found to have a mean diameter of 2.498 cm. with a SD of 0.012 cm. Determine the percentage of pieces produced you would except to lie within of the drawing limits of 2.5 ± 0.02 **cm.**

Task 4

An intelligence test is administered to 1000 children. The average score is 42 and S.D is 24. Assuming the test follows normal distribution

i) Find the number of children exceeding the score 60.

ii) Find the number of children with score lying between 20 and 40.

Task 5

The mean weight of 500 male students in a certain college is 151 *lb* and the standard deviation is 15*lb*. assuming the weights are normally distributed find how many students weight. (i) Between 142 and 155 *lb*. (ii) More than 185 *lb*.

Task 6

The saving bank account of a customer showed an average balance of Rs.1500 and a standard deviation of Rs.500 .assuming that the account balances are normally distributed.

(i) What percentage of account is over Rs.2000?

(ii) What percentage of account is between Rs.1200 and Rs.1700?

STEP 3: PRACTICE/TESTING

1. **What is the p.d.f. of a normal distribution?**

2. **Define standard normal distribution.**

3. **Mention some properties of normal distribution.**

FACULTY ASSESSMENT

Description	**Max Marks Awarded**
Preparation	10
Conduct of Experiment & Result	10
Viva	10
Total	30
Faculty Signature	

Experiment number : 5 **Date:**

APPLICATIONS OF STUDENT T-TEST

STEP 1: INTRODUCTION

OBJECTIVES OF THE EXPERIMENT

1. To apply t-test to test hypothesis about population mean
2. To apply t-test to test hypothesis about two means
3. To apply paired t-test to test hypotheses about means of two dependent samples

STEP 2: ACQUISITION

Student's t – distribution

Student's **t-distribution** has the probability density function given by

$$f(t) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\,\Gamma(\frac{\nu}{2})}\left(1+\frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}, \quad -\infty < t < \infty$$

where ν is the number of degrees of freedom and Γ is the gamma function. This may also be written as

$$f(t) = \frac{1}{\sqrt{\nu}\,B\left(\frac{1}{2},\frac{\nu}{2}\right)}\left(1+\frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}, \quad -\infty < t < \infty$$

Note: (a) The values of $t_\nu(\alpha)$ can be got from the t – table

(b) $t_\nu(2\alpha)$gives the critical value of t for a single tail test at α LOS and ν d.f

For eg, $t_8(0.05)$ for single tailed test = $t_8(10)$ for two-tailed test = 1.86

Test of Hypothesis about the Population Mean

Test statistic $t = \frac{\bar{x}-\mu}{S/\sqrt{n}}$ follows t – distribution with n-1 degrees of freedom.

where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2$

Null hypothesis H_0 : There is no significant difference between the sample mean $\bar{x}$ and the population mean μ.

If $|t| \leq$ tabulated t, then H_0 is accepted and the difference between $\bar{x}$ and μ is not considered significant.

Assumptions for t – test for population mean

1. The parent population from which the sample is drawn is normal.
2. The sample observations are independent
3. The population standard deviation σ is unknown.

Test of Hypothesis about the difference between two means

To test a hypothesis concerning the difference between the means of two normally distributed populations, when the population variances are unknown, t – test is used.

H_0: The samples have been drawn from populations with same means, ie, $\mu_1 = \mu_2$

Test statistic is $t = \dfrac{\bar{x} - \bar{y}}{S\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} \sim t_{n_1+n_2-2}$

where $\bar{x} = \dfrac{\sum x}{n_1}, \ \bar{y} = \dfrac{\sum y}{n_2}$,

$$S^2 = \frac{1}{n_1 + n_2 - 2}\left[\sum_i (x_i - \bar{x})^2 + \sum_j (y_j - \bar{y})^2\right]$$

or $$S^2 = \frac{1}{n_1 + n_2 - 2}\left[n_1 s_1^{\,2} + n_2 s_2^{\,2}\right]$$, where

$$s_1^{\,2} = \frac{1}{n_1}\sum_i (x_i - \bar{x})^2, \quad s_2^{\,2} = \frac{1}{n_2}\sum_j (y_j - \bar{y})^2$$

(Note : S^2 is an unbiased estimate of the population variance σ^2)

The test statistic follows t-distribution with $n_1 + n_2 - 2$ degrees of freedom.

If $|t| \leq$ tabulated t, then H_0 is accepted and the difference between $\bar{x}$ and μ is not considered significant.

Paired t-test for difference of Means

If the two given samples are dependent, ie, each observation in one sample is associated with a particular observation in the second sample, then we use paired t – test to test whether the means differ significantly or not. Here , both the samples will have same number of units.

The test statistic is

$$t = \frac{\bar{d}}{S/\sqrt{n}} \quad \text{where} \quad \bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i \quad \text{and} \quad d_i = x_i - y_i, \quad S^2 = \frac{1}{n-1}\sum (d_i - \bar{d})^2$$

t follows t – distribution with n-1 d.f. Here n is the number of pairs in the sample

Using R for testing of hypothesis

The R function t.test() can be used to perform both one and two sample t-tests on vectors of data. The function contains a variety of options and can be called as follows:

t.test(x, y = NULL, alternative = c("two.sided", "less", "greater"), mu = 0, paired = FALSE, var.equal = FALSE, conf.level = 0.95)

Here x is a numeric vector of data values and y is an optional numeric vector of data values. If y is excluded, the function performs a one-sample t-test on the data contained in x, if it is included it performs a two-sample t-tests using both x and y.

The option mu provides a number indicating the true value of the mean (or difference in means if you are performing a two sample test) under the null hypothesis. The option alternative is a character string specifying the alternative hypothesis, and must be one of the following: "two.sided" (which is the default), "greater" or "less" depending on whether the alternative hypothesis is that the mean is different than, greater than or less than mu, respectively.

Procedure for doing the Experiment:

1.	To test hypothesis about population mean: (a)For a two-tailed test x = c(a_1 , a_2 ,.....a_N) t.test(x,alternative="two.sided",mu= μ) (b) For a one-tailed test x = c(a_1 , a_2 ,.....a_N) t.test(x,alternative="less"/"greater",mu= μ)
2.	To test hypothesis about two means A= c(a_1 , a_2 ,.....a_m) B = c(b_1 , b_2 ,.....b_n) t.test(A,B,alternative="two.sided"/"less"/"greater",, var.equal=TRUE)
3.	To use paired t-test A= c(a_1 , a_2 ,.....a_m) B = c(b_1 , b_2 ,.....b_n) t.test(A,B,alternative="greater"/"less"/"two.sided",paired=TRUE)

EXAMPLE – Single mean

Eleven articles produced by a factory were chosen at random and their weights were found to be (in kgs) 63,63,66,67,68,69,70,70,71,71,71 respectively. In the light of the above data, can we assume that the mean weight of the articles produced by the factory is 66 kgs? (Given: the critical value of t for 10 degrees of freedom at 5% LOS is 2.28).

Null Hypothesis : $H_0 : \mu = 66$

Alternative Hypothesis : $H_1 : \mu \neq 66$

R-code

```
x = c(63,63,66,67,68,69,70,70,71,71,71)
t.test(x,alternative="two.sided",mu=66)
```

Output:

```
One Sample t-test
data:  x
t = 2.3, df = 10, p-value = 0.04425
alternative hypothesis: true mean is not equal to 66
95 percent confidence interval:
66.06533 70.11649
sample estimates:
mean of x
68.09091
```

Conclusion: t-value = 2.3 > 2.228. Hence we reject H_0 and we may conclude that the mean weight of the articles produced by the factory is not 66.

Task 1

Tests made on the breaking strength of 10 pieces of a metal gave the followingresults. 578, 572, 570, 568, 572, 570, 570, 572, 596 and 584 kg.
Test if the mean breaking strength of the wire can be assumed as 577kg.

Null hypothesis:

Alternate hypothesis:

R-code

Output:

Conclusion:

Task 2

The heights of 10 men in a given locality are found to be 70, 67, 62, 68, 61, 68, 70, 64, 64, 66 inches. Is it reasonable to believe that the average height is greater than 64 inches?

Null hypothesis H_0 :

Alternate hypothesis: H_1 **:**

R-code:

Output :

Conclusion:

Example 2: Two means

6 subjects were given a drug (treatment group) and an additional 6 subjects a placebo (control group). Their reaction time to a stimulus was measured (in ms).

Placebo group: 91, 87, 99, 77, 88, 91

Treatment group : 101, 110, 103, 93, 99, 104

Can we conclude that the reaction time of the placebo group is less than that of the treatment group? (Required table value of t = 1.1812)

Null hypothesis H_0 : $\mu_1 = \mu_2$, ie. the reaction times of the two groups are equal.

Alternate hypothesis H_1 : $\mu_1 < \mu_2$ ie, the reaction time of the placebo group is less than that of the treatment group

R-code:

```
Control = c(91, 87, 99, 77, 88, 91)
Treat = c(101, 110, 103, 93, 99, 104)
t.test(Control,Treat,alternative="less", var.equal=TRUE)
```

Output:

Two Sample t-test

data: Control and Treat t = -3.4456, df = 10, p-value = 0.003136 alternative hypothesis: true difference in means is less than 0

Conclusion: t-value =-3.4456 , $|t|$ =3.4456 > 1.1812. Hence we may conclude that the reaction time of placebo group is less than that of treatment group.

Task 3

Two independent samples are chosen from two schools A and B and common test is given in a subject. The scores of the students are as follows:

School A: 76 68 70 43 94 68 33

School B: 40 48 92 85 70 76 68 22.

Can we conclude that students of school A performed better than students of school B.

Null hypothesis

Alternate hypothesis

R-code:

Output:

Conclusion:

Task 4

Two independent samples of sizes 8 and 7 contained the following values.
Sample 1:19 17 15 21 16 18 16 14
Sample 2:15 14 15 19 15 18 16
Is the difference between the sample means significant?

Null hypothesis H_0 :

Alternate hypothesis H_1 :

R-code:

Output:

Conclusion:

Example 3: Paired t-test

A study was performed to test whether cars get better mileage on premium gas than on regular gas. Each of 10 cars was first filled with either regular or premium gas, decided by a coin toss, and the mileage for that tank was recorded. The mileage was recorded again for the same cars using the other kind of gasoline. The relevant mileages : Regular: 16, 20, 21, 22, 23, 22, 27, 25, 27, 28 Premium :19, 22, 24, 24, 25, 25, 26, 26, 28, 32 . Use a paired t test to determine whether cars get significantly better mileage with premium gas.

Null Hypothesis H_0 : $\mu_1 = \mu_2$, ie, the two types of bulbs are identical regarding length of life.

Alternative Hypothesis: $H_1 : \mu_2 > \mu_1$

```
reg=c(16,20,21,22,23,22,27,25,27,28)

prem=c(19,22,24,24,25,25,26,26,28,32)

t.test(prem,reg,alternative="greater",paired=TRUE)
```

Paired t-test

data: prem and reg

t = 4.4721, df = 9, p-value = 0.0007749

alternative hypothesis: true difference in means is greater than 0

95 percent confidence interval:

1.180207 Inf

sample estimates:

mean of the differences

2

Conclusion: p-value = 0.0007749 < 0.05 Hence we reject H_0 and we may conclude that cars get significantly better mileage with premium gas.

Task 5

The weight gain in pounds under two systems of feeding of calves of 10 pairs of identical twins is given below.

Twin pair	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Weight gain under System A	**43**	**39**	**39**	**42**	**46**	**43**	**38**	**44**	**51**	**43**
Weight gain under System B	**37**	**35**	**34**	**41**	**39**	**37**	**37**	**40**	**48**	**36**

Discuss whether the difference between the two systems of feeding is significant.

Null Hypothesis H_0 :

Alternative Hypothesis: H_1 :

R-code:

Output:

Conclusion:

Task 6

Ten persons were appointed in the officer cadre in an office. Their performance was noted by giving a test and the marks were recorded out of 100.

Employee	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**J**
Before training	**80**	**76**	**92**	**60**	**70**	**56**	**74**	**56**	**70**	**56**
After training	**84**	**70**	**96**	**80**	**70**	**52**	**84**	**72**	**72**	**50**

By applying t test, can it be concluded that the employees have been benefited by the training?

Null hypothesis:

Alternate hypothesis:

R-code:

Output:

Conclusion:

STEP 3: PRACTICE/TESTING

1. **Write the test statistic for testing hypothesis about a population mean.**

2. **Write the test statistic for testing of hypothesis about the difference between two means .**

3. **Write the test statistic for testing of hypothesis about the difference between means of two dependent samples. (paired t-test)**

4. **Define level of significance.**

FACULTY ASSESSMENT

Description	**Max Marks Awarded**
Preparation	10
Conduct of Experiment & Result	10
Viva	10
Total	30
Faculty Signature	

Experiment number : 6 **Date:**

APPLICATIONS OF F TEST

STEP 1: INTRODUCTION

OBJECTIVES OF THE EXPERIMENT

To apply F-test to compare the variances of two samples from normal populations.

STEP 2: ACQUISITION

The null hypothesis is that the ratio of the variances of the populations from which x and y were drawn, or in the data to which the linear models x and y were fitted, is equal to ratio.

Procedure for doing the Experiment:

	R-Code for F-test: var.test(x, y, ratio = 1,alternative = c("two.sided", "less", "greater"),conf.level = 0.95, ...)

Note:

x, y	-	numeric vectors of data values, or fitted linear model objects (inheriting from class "lm").
Ratio	-	the hypothesized ratio of the population variances of x and y.
Alternative	-	a character string specifying the alternative hypothesis, must be one of "two.sided" (default), "greater" or "less". You can specify just the initial letter.
conf.level	-	confidence level for the returned confidence interval.

In the test statistic, the greater of the two variances ${S_1}^2$ and ${S_2}^2$ is to be taken in the numerator and v_1 corresponds to the greater variance.

Example:

Two samples of 6 and 7 items respectively have the following values for a variable

Sample 1	**39**	**41**	**42**	**42**	**44**	**40**	
Sample 2	**40**	**42**	**39**	**45**	**38**	**39**	**40**

Do the sample variances differ significantly?

Null Hypothesis: There is no significant difference in sample variances.

Alternative Hypothesis: There is a significant difference in sample variances.

Code:

```
x=c(40,42,39,45,38,39,40)
y=c(39,41,42,42,44,40)
var.test(x, y, ratio = 1,
alternative = c("two.sided"),
conf.level = 0.95)
```

Output:

```
F test to compare two variances
data:  x and y
F = 1.8323, numdf = 6, denomdf = 5, p-value = 0.523
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.2625934 10.9710044
sample estimates:
ratio of variances
       1.832298
```

Critical value of F for (6, 5) d.f. is $F_{0.05} = 4.95$

Conclusion:Since $F < F_{0.05}$, weaccept the null hypothesis and we may conclude that there is no significant difference in the sample variances.

Task 1:

Two random samples drawn from two normal populations are

Sample 1:	**20**	**16**	**26**	**27**	**23**	**22**	**18**	**24**	**25**	**19**	
Sample 2:	**27**	**33**	**42**	**35**	**32**	**34**	**38**	**28**	**41**	**43**	**30**	**37**

Test whether the populations have the same variances.

Null Hypothesis:

Alternative Hypothesis:

R Code:

Output:

Conclusion:

Task 2:

The nicotine content in 2 random samples of tobacco are given below:

Sample 1:	**21**	**24**	**25**	**26**	**27**	
Sample 2:	**22**	**27**	**28**	**30**	**31**	**36**

Test whether the populations have the same variances.

Null Hypothesis:

Alternative Hypothesis:

R Code:

Output:

Conclusion:

Task 3:

2 independent samples of 8 and 7 items have the following values.

Sample 1:	**9**	**11**	**13**	**11**	**15**	**9**	**12**	**14**
Sample 2:	**10**	**12**	**10**	**14**	**9**	**8**	**10**	

Can we conclude that the two samples have drawn from the same normal population.

To test whether the samples come from the same normal population, we have to test for

a. Equality of population means
b. Equality of population variances.

Equality of means is tested using t-test and equality of variances is tested using F-test.

Since t-test assumes $\sigma_1^2 = \sigma_2^2$, we first apply *F*-test and then t-test.

***F*-test:**

Null Hypothesis:

Alternative Hypothesis:

R Code:

Output:

Conclusion:

***t*-test:**

Null Hypothesis:

Alternative Hypothesis:

R Code:

Output:

Conclusion:

Final conclusion:

Task 4:

Two horses A and B were tested according to the time(in seconds) to run a particular track with the following results:

Horse A: 28 30 32 33 33 29 34

Horse B: 29 30 30 24 27 29

Test whether the two horses have the same running capacity in terms of average and variance of time taken.

Null Hypothesis:

Alternative Hypothesis:

R Code:

Output:

Conclusion:

Task 5:

Two samples are drawn from two normal populations. From the following data 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.**

Null Hypothesis:

Alternative Hypothesis:

R Code:

Output:

Conclusion:

STEP 3: PRACTICE/TESTING

1. **What is the use of *F*-distribution?**

2. **State the important properties of *F*-distribution.**

3. **What is the difference between *F*-test and *t*-test?**

FACULTY ASSESSMENT

Description	**Max Marks Awarded**
Preparation	10
Conduct of Experiment & Result	10
Viva	10
Total	30
Faculty Signature	

Experiment number : 7 **Date:**

APPLICATION OF CHI SQUARE TEST

STEP 1: INTRODUCTION

OBJECTIVES OF THE EXPERIMENT

1. To apply chi square test for goodness of fit
2. To apply chi square test for independence of attributes

STEP 2: ACQUISITION

Conditions for the validity of χ^2 -test

1. The sample observations must be independent of one another.
2. The sample size must be reasonably large, say $\geq$ 50.
3. No individual frequency should be less than 5. If any frequency is less than 5, then it is pooled with the preceding or succeeding frequency so that the pooled frequency is more than 5. Finally adjust for the d.f lost in pooling.
4. The number of classes k must be neither too small nor too large, ie $4 \leq k \leq 16$

χ^2 -test of goodness of fit

Tests ofgoodness of fit are used when we want to determine whether an actual sample distribution matches a known theoretical distribution. It enables us to find if the deviation of the experiment from theory is just by chance or it is due to the inadequacy of the theory to fit the data.

Null Hypothesis: H_0: The difference between the observed and expected frequencies is not significant. ie, the theory fits well into the given data.

Regular method: Let $O_i (i = 1,2,....n)$ be a set of observed frequencies and $E_i (i = 1,2,....n)$ be the corresponding set of expected frequencies. Then $\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$ follows Chi- Square Distribution with n – 1 d.f.

(One degree of freedom is subtracted for the constraint $\sum_i O_i = \sum_i E_i$)

Compare the calculated χ^2 -value with the tabulated χ^2 -value (with n – 1 d.f) and form the conclustion.

χ^2 - test of Independence of Attributes

χ^2 - test isused for testing the null hypothesis that two criteria of classification are independent. Let the two attributes be *A* and *B*, where *A* has *r* categories and *B* has *s* categories. Thus the members of the population and hence, those of the sample are divided into rs classes. Let the total number of observations be *N*. The observations are arranged in the form of a matrix, called contingency table .

H_0:The attributes *A* and *B* are independent.

Regular method:

The expected frequencies E_{ij} for various cells are calculated using the formula:

$$E_{ij} = \frac{R_i C_j}{N} \ , \ i = 1,2,....r, \quad j = 1,2.....s$$

$$= \frac{\text{Total of observed frequencies in the i}^{\text{th}} \text{ row} \times \text{Total of observed frequencies in the j}^{\text{th}} \text{column}}{\text{Total frequency}}$$

Test statistic is $\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}}$ which follows χ^2 - distribution with $n = (r-1)(s-1)$ degrees of freedom.

Note: For a 2×2 contingency table with cell frequencies a, b, c, d , the χ^2 - value is given by

$$\chi^2 = \frac{N(ad - bc)^2}{(a+c)(b+d)(a+b)(c+d)}; \qquad N = a + b + c + d \ ,$$

Degree of freedom =1

Procedure for doing the Experiment:

1.	**R-code for testing goodness of fit:** f=vector of observed frequencies p= vector of expected ratios (probabilities) a=chisq.test(f,p=c(p_1 ,p_2 , ….)) a
2.	**R-code for testing independence of attributes:** a = vector of elements in first row of contingency table b = vector of elements in second row of contingency table c = ……. contingency = as.data.frame(rbind(a,b,c,….)) # to create the table

	contingency chisq.test(contingency,simulate.p.value=T)

Example: (χ^2 -test of goodness of fit)

The following table gives the number of aircraft accidents that occur during the various days of a week. Find whether the accidents are uniformly distributed over the week.

Days	**Sun**	**Mon**	**Tue**	**Wed**	**Thu**	**Fri**	**Sat**
No. of accidents:	**14**	**16**	**8**	**12**	**11**	**9**	**14**

Null Hypothesis: The accidents are uniformly distributed over the week

Alternative Hypothesis: The accidents are not uniformly distributed over the week

Level of significance: 5% (say)

R-code:

```
accident=c(14,16,8,12,11,9,14)
p=c(1/7,1/7,1/7,1/7,1/7,1/7,1/7)
a=chisq.test(accident,p=c(1/7,1/7,1/7,1/7,1/7,1/7,1/7))
a
```

Output:

Chi-squared test for given probabilities

data: accident

X-squared = 4.1667, df = 6, p-value = 0.6541

Table value of $\chi^2_{0.05}$ for 6 d.f = 12.59

Conclusion: $\chi^2 < \chi^2_{0.05}$,so we accept H_0 and conclude that the accidents are uniformly distributed over the week.

(Or)

Here *p* value $\geq \alpha$ value,so we accept H_0 and conclude that the accidents are uniformly distributed over the week.

Task 1

The following figures show the distribution of digits in numbers chosen at random from a telephone directory

Digits	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**Total**

Frequency	**1026**	**1107**	**997**	**966**	**1075**	**933**	**1107**	**972**	**964**	**853**	**10000**

Test whether the digits may be taken to occur equally frequently in the directory.

Null Hypothesis:

Alternative Hypothesis:

Level of significance:

R-code:

Output:

Table value of $\chi^2_{0.05}$ for 6 d.f =

Conclusion:

Task 2

The following is the distribution of the hourly number of trucks arriving at a company's warehouse:

Trucks arriving hour	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**Total**
Frequency	**52**	**51**	**56**	**47**	**60**	**57**	**59**	**61**	**57**	**500**

Test whether the arrival of trucks is equally distributed at the 0.05 level of significance.

Example (χ^2 - test of Independence of Attributes)

A survey of 920 people that ask for their preference of one of three ice cream flavours (chocolate, vanilla, strawberry) gives the following results:

	Flavour				
Gender		**Chocolate**	**Vanilla**	**Strawberry**	**Total**
	Men	**100**	**120**	**60**	**280**
	Women	**350**	**200**	**90**	**640**
	Total	**450**	**320**	**150**	**920**

Using χ^2 test, determine whether or not there is an association between gender and preference for ice cream flavour.

R-code

```
men = c(100, 120, 60)
```

women = c(350, 200, 90)

icecream = as.data.frame(rbind(men, women))

chisq.test(icecream,simulate.p.value=T)

Output:

V1 V2 V3

men 100 120 60

women 350 200 90

>chisq.test(icecream, simulate.p.value=T)

Pearson's Chi-squared test with simulated p-value (based on 2000 replicates)

data: icecream

X-squared = 28.362, df = NA, p-value = 0.0004998

Table value of χ^2 =5.991

Conclusion: $\chi^2 > \chi_\alpha^2$, hence we conclude that there is association between gender and preference for ice cream flavour.

Note:

The R-code

men = c(100, 120, 60)

women = c(350, 200, 90)

ice.cream.survey = as.data.frame(rbind(men, women))

ice.cream.survey

generates the table

V1 V2 V3

men 100 120 60

women 350 200 90

Task 3

Two sample polls of votes for two candidates A and B for a public office are taken, one from among the residents of rural areas and one from the residents of urban areas. The results are given in the table. Examine whether the nature of the area is related to voting preference in the election

Votes for area	A	B	Total
Rural	620	380	1000
Urban	550	450	1000
Total	1170	830	2000

Null Hypothesis: H_0:

Alternative Hypothesis: H_1:

Level of significance: α =

R-code:

Output:

Table value:

Conclusion:

Task 4

A sample of 200 persons with a particular disease was selected. Out of these, 100 were given a drug and the others were not given any drug. The results are as follows:

No. of persons	Drug	No drug
Cured	65	55
Not cured	35	45

Test whether the drug is effective or not (Use α = 0.05)

Null Hypothesis: H_0:

Alternative Hypothesis: H_1:

Level of significance: α =

R-code:

Output:

Table value:

Conclusion:

Task 5

The following data are collected on two characters.

	Smokers	**Non – Smokers**
Literates	**83**	**57**
Illiterates	**45**	**68**

Based on this, can you say that there is no relation between smoking and literacy?

Null Hypothesis: H_0:

Alternative Hypothesis: H_1:

Level of significance: α =

R-code:

Output:

Table value:

Conclusion:

Task 6

From the following data, test whether there is any association between intelligence and economic conditions?

	Intelligence				
Economic condition	**Excellent**	**Good**	**Medium**	**Dull**	**Total**
Good	**48**	**200**	**150**	**80**	**478**
Not good	**52**	**180**	**190**	**100**	**522**
Total	**100**	**380**	**340**	**180**	**1000**

Null Hypothesis: H_0:

Alternative Hypothesis: H_1:

Level of significance: α =

R-code:

Output:

Table value:

Conclusion:

STEP 3: PRACTICE/TESTING

1. **When is chi-square test used?**

2. **State the conditions for the validity of χ^2 -test**

3. **When do we use χ^2 -test of goodness of fit ?**

4. **When do we use χ^2 - test of Independence of Attributes?**

FACULTY ASSESSMENT

Description	**Max Marks Awarded**
Preparation	10
Conduct of Experiment & Result	10
Viva	10
Total	30
Faculty Signature	

Experiment number : 8 **Date:**

ANOVA – ONE WAY CLASSIFICATION

STEP 1: INTRODUCTION

OBJECTIVES OF THE EXPERIMENT

To perform analysis of variance for a completely randomized design

STEP 2: ACQUISITION

Analysis of variance refers to the separation of variance ascribable to one group of causes from the variance ascribable to the other group. It is used to test the homogeneity of several means.

Three types of variation are present in a data

1. Treatments
2. Environmental
3. Residual or Error

Assumptions for ANOVA test

1. The observations are independent.
2. The parent population is normal
3. Various treatment and environmental effects are additive in nature.
4. The samples have been randomly selected from the population

Null Hypothesis: All the population means are equal

Alternative Hypothesis: Some of the means are not equal.

Three important designs of experiments:

1. Completely Randomised Design (CRD) – One-way classification
2. Randomised Block Design (RBD) – Two-way classification
3. Latin Square Design (LSD) – Three-way classification

Procedure for doing the Experiment:

1.	aov(response~factor,data=data_name)

Example

A drug company tested three formulations of a pain relief medicine for migraine headac hesufferers. For the experiment 27 volunteers were selected and 9 were randomly assign ed to one of three drug formulations. The subjects were instructed to take the drug duri ng theirnext migraine headache episode and to report their pain on a scale of 1 to 10 (10 beingmaximum pain)

Drug A	**4**	**5**	**4**	**3**	**2**	**4**	**3**	**4**	**4**
Drug B	**6**	**8**	**4**	**5**	**4**	**6**	**5**	**8**	**6**
Drug C	**6**	**7**	**6**	**6**	**7**	**5**	**6**	**5**	**5**

R-code:

```
pain=c(4,5,4,3,2,4,3,4,4,6,8,4,5,4,6,5,8,6,6,7,6,6,7,5,6,5,5)

drug=c(rep("A",9),rep("B",9),rep("C",9))

data=data.frame(pain,drug)

data

results=aov(pain~drug,data=data)
summary(results)
```

Output:

```
pain drug

1    4    A
2    5    A
3    4    A
4    3    A
5    2    A
6    4    A
7    3    A
8    4    A
9    4    A
10   6    B
11   8    B
12   4    B
13   5    B
14   4    B
15   6    B
16   5    B
17   8    B
18   6    B
19   6    C
20   7    C
21   6    C
22   6    C
23   7    C
24   5    C
25   6    C
26   5    C
27   5    C
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
drug	2	28.22	14.111	11.91	0.000256 ***
Residuals	24	28.44	1.185		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

$F_\alpha = 3.40, F > F_\alpha$, so we reject the null hypothesis and conclude that that the means of the three drug groups are different.

Task 1

Three machines A, B & C gave the production of pieces in 4 days as below is there a significant difference between machines?

A	**17**	**16**	**14**	**13**
B	**15**	**12**	**19**	**18**
C	**20**	**8**	**11**	**17**

Task 2

Four machines A,B,C,D are used to produce a certain kind of cotton fabric. 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 is counted with 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 significant difference in the performance of the four machines?

Task 3

Ten varieties of wheat are grown in 3 plots each and the following yields in quintals per acre, obtained.

		Variety									
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Plots	**I**	**7**	**7**	**14**	**11**	**9**	**6**	**9**	**8**	**12**	**9**
	II	**8**	**9**	**13**	**10**	**9**	**7**	**13**	**13**	**11**	**11**
	III	**7**	**6**	**16**	**11**	**12**	**6**	**12**	**11**	**11**	**11**

Test the significance of the differences between variety yields

Task 4

An experiment was conducted to study effect of four different dyes A, B, C, D on the strength of the fabric and following results of fabric strength are obtained.

Dye					
A	**8.67**	**8.68**	**8.66**	**8.65**	
B	**7.68**	**7.58**	**8.67**	**8.65**	**8.62**
C	**8.69**	**8.67**	**8.92**	**7.7**	
D	**7.7**	**7.90**	**8.65**	**8.20**	**8.60**

STEP 3: PRACTICE/TESTING

1. **What are the basic principles of Experimental Design?**

2. **Mention the important designs of experiments:**

3. **Explain a completely randomized design.**

4. **What is the purpose of analysis of variance?**

FACULTY ASSESSMENT

Description	**Max Marks Awarded**
Preparation	10
Conduct of Experiment & Result	10
Viva	10
Total	30
Faculty Signature	

Experiment number : 9 **Date:**

ANOVA – TWO WAY CLASSIFICATION

STEP 1: INTRODUCTION

OBJECTIVES OF THE EXPERIMENT

To perform analysis of variance for a Randomised Block Design.

STEP 2: ACQUISITION

The data collected from experiments with randomised block design form a two-way classification, classified according to two factors – blocks and treatments. The two-way table has k rows and r columns – ie, N=kr entries.

Consider an agricultural experiment in which we wish to test the effect of k fertilising treatments on the yield of a crop. We divide the plots into r blocks, according to soil fertility, each block containing k plots. The plots in each block will be of homogeneous fertility. I each block, the k treatments are given to the k plots in a random manner in such a way that each treatment occurs only once in each block. The same k treatments are repeated from block to block.

H_{01} : There is no difference in the yield of crop due to treatments

H_{02} : There is no difference in the yield of crop due to blocks

Procedure for doing the Experiment:

Consider a two way table with k rows and r columns

1.	a=c(a_1 , a_2 ,……) (entries entered columnwise) f=c("row1","row2","row3","row4","row5") k=5 r=4 A=gl(k,1,r*k,factor(f)) A B=gl(r,k,k*r) B av = aov(a ~ A+B) summary(av)

Example

The following data represents the number of units of loom crank bushes produced per day turned out by different workers using four different types of machines.

		Machine Type			
		A	**B**	**C**	**D**
	1	**44**	**38**	**47**	**36**
Workers	**2**	**46**	**40**	**52**	**43**
	3	**34**	**36**	**44**	**32**
	4	**43**	**38**	**46**	**33**
	5	**38**	**42**	**49**	**39**

Test whether the 5 men differ with respect to mean productivity and test whether the mean Productivity is the same for the four different machine types.

R-code:

```
a=c(44,46,34,43,38,38,40,36,38,42,47,52,44,46,49,36,43,32,33,39)
f=c("w1","w2","w3","w4","w5")
k=5
r=4
worker=gl(k,1,r*k,factor(f))
worker
machine=gl(r,k,k*r)
machine
av = aov(a ~ worker+machine)
summary(av)
```

Output:

```
a=c(44,46,34,43,38,38,40,36,38,42,47,52,44,46,49,36,43,32,33,39)
 f=c("w1","w2","w3","w4","w5")
 k=5
 r=4
worker=gl(k,1,r*k,factor(f))
worker
 [1] w1 w2 w3 w4 w5 w1 w2 w3 w4 w5 w1 w2 w3 w4 w5 w1 w2 w3 w4 w5
Levels: w1 w2 w3 w4 w5
machine=gl(r,k,k*r)
machine
```

[1] 1 1 1 1 1 2 2 2 2 2 3 3 3 3 3 4 4 4 4 4

Levels: 1 2 3 4

>av = aov(a ~ worker+machine)

>summary(av)

Df Sum Sq Mean Sq F value Pr(>F)

worker 4 161.5 40.37 6.574 0.00485 **

machine 3 338.8 112.93 18.388 8.78e-05 ***

Residuals 12 73.7 6.14

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Conclusion:

From F-table, $F_{0.05}(4,12) = 3.26$

$F_{0.05}(3,12) = 3.49$

$F_1 = 6.54 > F_{0.05}(4,12) = 3.26$, hence we reject H_{01} and conclude that the 5 workers differ with respect to mean productivity.

$F_2 = 18.388 > F_{0.05}(3,12) = 3.49$, hence we reject H_{02} and conclude that the 4 machines differ with respect to mean productivity.

Task 1

A company appoints 4 salesmen A,B,C,D and observes their sales in 3 seasons: summer, winter and monsoon. The figures (in lakhs of Rs.) are given in the following table:

	Salesmen			
Season	**A**	**B**	**C**	**D**
Summer	**45**	**40**	**38**	**37**
Winter	**43**	**41**	**45**	**38**
Monsoon	**39**	**39**	**41**	**41**

Carry out an analysis of variance.

Task 2

Four different, though supposedly equivalent, forms of a standardized reading achievement test were given to each of 5 students and the following are the scores which they obtained:

	Student 1	Student 2	Student 3	Student 4	Student 5
Form A	**75**	**73**	**59**	**69**	**84**
Form B	**83**	**72**	**56**	**70**	**92**
Form C	**86**	**61**	**53**	**72**	**88**
Form D	**73**	**67**	**62**	**79**	**95**

Perform a two-way analysis of variance to test at the level of significance 0.01 whether it is reasonable to treat the forms as equivalent.

Task 3

An experiment was designed to study the performance of different detergents for cleaning fuel injectors. The following 'cleanness' readings were obtained with specially designed equipment's for 12 tanks of gas distributed over 3 different models of engines:

	Engine 1	Engine 2	Engine 3	Total
Detergent A	**45**	**43**	**51**	**139**
Detergent B	**47**	**46**	**52**	**145**
Detergent C	**48**	**50**	**55**	**153**
Detergent D	**42**	**37**	**49**	**128**
Total	**182**	**176**	**207**	**565**

Test at the 0.01 level of significance whether there are differences in the detergents or in the engines.

Task 4:

Four experiments determine the moisture content of samples of a powder each observer taking a sample from each of six consignments. The assessments are given below

Observer	Consignment					
	1	2	3	4	5	6
1	9	10	9	10	11	11
2	12	11	9	11	10	10
3	11	10	10	12	11	10
4	12	13	11	14	12	10

Perform an analysis of variance on these data and discuss whether there is any significant difference between consignments or between observers.

STEP 3: PRACTICE/TESTING

1. **What is meant by a randomized block design?**

2. **Write the differences between CRD and RBD.**

3. **Bring out any two advantages of RBD over CRD.**

4. **When do you apply the analysis of variance technique?**

FACULTY ASSESSMENT

Description	**Max Marks Awarded**
Preparation	10
Conduct of Experiment & Result	10
Viva	10
Total	30
Faculty Signature	

Experiment number : 10 **Date:**

CONTROL CHARTS FOR VARIABLES (MEAN AND RANGE CHART)

STEP 1: INTRODUCTION

OBJECTIVES OF THE EXPERIMENT

To plot $\overline{X}$ - chart and R-chart and comment on the state of control of the process.

STEP 2: ACQUISITION

Statistical Quality Control is a statistical method for finding whether the variation in the quality of the product is due to random causes or assignable causes

Control chart is a graphical device used in statistical quality control for the study and control of the manufacturing process.

There are two types of control charts:

1. Control charts of variables (Mean ($\overline{X}$)and range (R)charts)
2. Control charts of attributes (p-chart, np-chart, c-chart)

The Lower control limit and Upper control limit for mean and range charts

1. $\overline{X}$ chart LCL: $\overline{\overline{X}} - A_2\overline{R}$ UCL: $\overline{\overline{X}} + A_2\overline{R}$
2. R-Chart LCL: $D_3\overline{R}$ UCL: $D_4\overline{R}$

Procedure to plot $\overline{X}$ and R charts using RStudio

To install qcc package in RStudio go to the "Tools" menu, select "Install Packages…" and type "qcc" into the packages field being sure to also select "Install Dependencies" and click "Install."

Load the data from a.csv file with one subgroup per row :

my.data = read.csv("my-data.csv",header=FALSE)
OR,
Load the data for each subgroup manually:
a1 = c( )
a2 = c( )
a3 = c( ) etc.
If there is more than one subgroup, create a dataframe: my.data = rbind(a1,a2,a3)

Procedure for doing the Experiment:

Suppose the given values are x, y, z, …….

1.	**R code to create dataframe** S1=c(a_1 , a_2 ,……) S2=c(b_1 , b_2 ,……) A= as.data.frame(rbind(S1,S2,……..)) A

2.	**For $\overline{X}$ chart:**
	
	Xbarchart= qcc(data = A,
	type = "xbar",
	sizes = n, # n=number of items in each sample
	title = "X-bar Chart ",
	plot = TRUE)

3.	**For R chart:**
	
	rchart = qcc(data = A,
	type = "R",
	sizes = n, # n=number of items in each sample
	title = "R Chart",
	plot = TRUE)

Example

The measurements are given below with 5 samples each containing 5 items at equal intervals of time. Construct $\overline{X}$ and R charts and comment on the state of control.

Sample no	Measurements				
1	**46**	**45**	**44**	**43**	**42**
2	**41**	**41**	**44**	**42**	**40**
3	**40**	**40**	**42**	**40**	**42**
4	**42**	**43**	**43**	**42**	**45**
5	**43**	**44**	**47**	**47**	**45**

#R code to create dataframe

S1=c(46,45,44,43,42)

S2=c(41,41,44,42,40)

S3=c(40,40,42,40,42)

S4=c(42,43,43,42,45)

S5=c(43,44,47,47,45)

A= as.data.frame(rbind(S1,S2,S3,S4,S5))

A

#For $\overline{X}$ chart:

Xbarchart= qcc(data = A,

type = "xbar",

sizes = 5,

title = "X-bar Chart ",

plot = TRUE)

Output:

V1 V2 V3 V4 V5

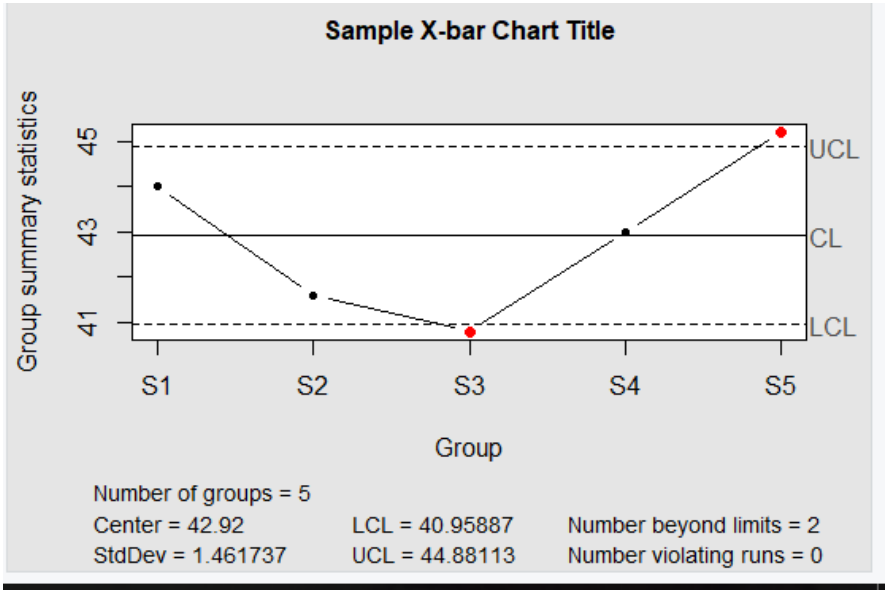
S1 46 45 44 43 42

S2 41 41 44 42 40

S3 40 40 42 40 42

S4 42 43 43 42 45

S5 43 44 47 47 45



For R chart:

R-code:

rchart = qcc(data = A,

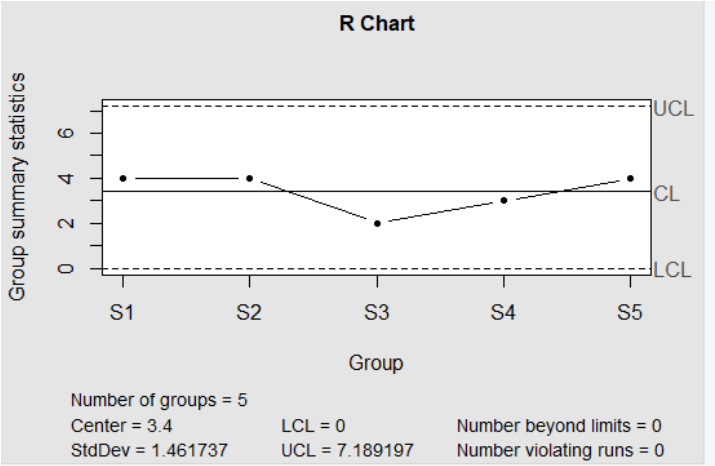
type = "R",

sizes = 5,

title = "R Chart",

plot = TRUE)

Output:



Conclusion:

In $\overline{X}$ chart, two points are beyond the control limits, so as far as sample mean is concerned, the system is out of control.

In R chart, all points are within the control limits, so as far as variability is concerned, the system is under control.

On the whole, the system is out of control.

Task 1

Samples of five ring bobbins each selected from a ring frame for eight shifts have shown following results of count of yarn.

Sample no.	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Count of yarn	**27.5**	**27.4**	**25.4**	**28.5**	**28.5**	**28.9**	**28.0**	**28.4**
	28.5	**26.9**	**26.9**	**28.0**	**29.0**	**29.5**	**28.5**	**28.5**
	28	**26.0**	**28.0**	**29.2**	**28.5**	**30.0**	**27.8**	**28.4**
	26.9	**28.7**	**26.7**	**29.0**	**28.5**	**29.4**	**28.0**	**28.0**
	28.6	**29.0**	**28.2**	**28.7**	**28.0**	**28.9**	**28.1**	**28.7**

Draw $\overline{X}$ and R chart for the above data and write conclusion about the state of the process.

Task 2:

The following data gives the measurements of 10 samples each of size 5, in a production process taken at intervals of 2 hours. Draw the control charts for the mean and range and comment on the state of control:

Sample No.	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Measurements	**47**	**52**	**48**	**49**	**50**	**55**	**50**	**54**	**49**	**53**
	49	**55**	**53**	**49**	**53**	**55**	**51**	**54**	**55**	**50**
	50	**47**	**51**	**49**	**48**	**50**	**53**	**52**	**54**	**54**
	44	**56**	**50**	**53**	**52**	**53**	**46**	**54**	**49**	**47**
	45	**50**	**53**	**45**	**47**	**57**	**50**	**56**	**53**	**51**

Task 3:

Plot the mean and range charts for the following data

	Rotation Time (msec)					
Sample Number	**1**	**2**	**3**	**4**	**5**	**6**
1	**469.92**	**468.67**	**479.76**	**454.38**	**469.58**	**454.46**
2	**457.34**	**454.37**	**475.28**	**453.46**	**480.03**	**480.40**
3	**473.96**	**459.26**	**460.42**	**462.04**	**450.60**	**451.52**
4	**480.06**	**469.86**	**456.42**	**460.63**	**465.66**	**466.99**
5	**467.46**	**476.56**	**474.01**	**465.34**	**475.27**	**462.97**
6	**473.06**	**475.86**	**472.97**	**454.93**	**470.73**	**466.24**
7	**456.27**	**476.37**	**479.50**	**459.86**	**470.73**	**452.35**

STEP 3: PRACTICE/TESTING

1. **Define Statistical Quality Control.**

2. **What are control charts? What are the types of control charts?**

3. **Write the Lower control limit and Upper control limit for mean and range charts.**

FACULTY ASSESSMENT

Description	**Max Marks Awarded**
Preparation	10
Conduct of Experiment & Result	10
Viva	10
Total	30
Faculty Signature	