KUMARAGURU COLLEGE OF TECHNOLOGY

LABORATORY MANUAL

Experiment Number: 4

Lab Code	**: U18MAI4201**
Lab	**: Probability and Statistics**
Course / Branch	**: B.E-CSE,ISE, B.Tech-IT**
Title of the Experiment	**: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.

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

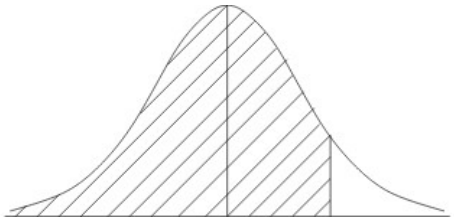
If a random variable X follows the normal distribution, then we write: $X \sim N(\mu, \sigma^2)$
The normal distribution with μ = 0 and σ = 1 is called the standard normal distribution, and is denoted as N(0,1).

Consider a normal distribution with mean μ and standard deviation σ

R-code for doing the Experiment:

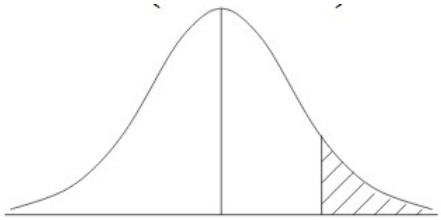
1.	To find $P(X < a) = P(-\infty < X < a)$ **R-code :** pnorm(a, mean = μ, sd = σ)
2.	To find $P(X > a) = P(a < X < \infty)$ R-code: pnorm(a, mean = μ, sd = σ, lower.tail = FALSE)
3.	To find $P(a < X < b)$ R-code: pnorm(b, mean = μ, sd = σ) - pnorm(a, mean = μ, sd = σ)

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



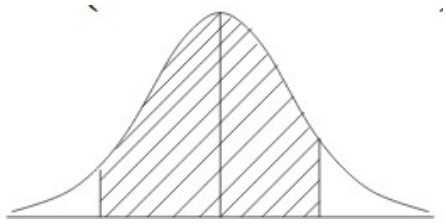
pnorm (a, mean = μ, sd = σ)

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)**

Ans:

(i) pnorm(71, mean=68,sd=2.5)-pnorm(66,mean=68,sd=2.5)

[1] 0.6730749

Percentage = 67.30749%

(ii) pnorm(72.5,mean=68,sd=2.5)-pnorm(71.5,mean=68,sd=2.5)

[1] 0.04482634

Percentage = 4.482634%

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)

Ans:

(i) over=pnorm(700,mean=662,sd=32,lower.tail=FALSE)

print(over*1000)

[1] 117.5152

(ii) below=pnorm(650,mean=662,sd=32)

print(below*1000)

[1] 353.8302

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.**

Ans:

pnorm(2.52,mean=2.498,sd=0.012)-pnorm(2.48,mean=2.498,sd=0.012)
[1] 0.8998163
Percentage = 89.998163%

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.**

Ans:

(i) pnorm(60,mean=42,sd=24,lower.tail=FALSE)
[1] 0.2266274

Number of children exceeding the score 60 = 0.2266274*1000 = 226.6 ≈ 227

(ii) pnorm(40,mean=42,sd=24)-pnorm(20,mean=42,sd=24)
[1] 0.2871346

Number of children with score lying between 20 and 40 = 0.2871346*1000
= 287.1346 ≈ 287

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*.

Ans:

(i)
```
one=pnorm(155,mean=151,sd=15)-pnorm(142,mean=151,sd=15)
print(one*500)
```

```
[1] 165.442
```

Number of students weigh between 142 and 155 lb≈ 166

(ii)
```
two=pnorm(185,mean=151,sd=15,lower.tail=FALSE)
print(two*500)
```

```
[1] 5.852649
```

Number of students weigh more than 185 lb≈ 6

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?

Ans:

(i)
```
pnorm(2000,mean=1500,sd=500,lower.tail=FALSE)
[1] 0.1586553
```

Percentage = 0.1586553*100 = **15.86 %**

(ii)
```
pnorm(1700,mean=1500,sd=500)-pnorm(1200,mean=1500,sd=500)
[1] 0.3811686
```

Percentage = 0.3811686*100 = **38.12 %**

STEP 3: PRACTICE/TESTING

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

A continuous random variable X follows normal distribution (or Gaussian distribution) then its p.d.f is

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \quad , -\infty < x < \infty, \quad -\infty < \mu < \infty, \quad \sigma > 0$$

The parameters are μ and σ where μ is the mean and σ is the standard deviation of thedistribution.

2. **Define standard normal distribution.**

The Normal Distribution with mean = 0 and variance = 1. If X is a normal variate, then $z = \frac{X-\mu}{\sigma}$ is a standard normal variate. The p.d.f of the standard normal variate is

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

Area under the standard normal curve = 1

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

1. The graph of the distribution is bell shaped and is called the normal probabilitycurve.
2. The curve is symmetrical about the ordinate at x = μ.
3. x – axis is an asymptote to the curve.
4. For the normal distribution, mean = median =mode.

KUMARAGURU COLLEGE OF TECHNOLOGY

LABORATORY MANUAL

Experiment Number: 2

Lab Code	**: U18MAI4201**
Lab	**: Probability and Statistics**
Course / Branch	**: B.E-CSE,ISE, B.Tech-IT**
Title of the Experiment	**: 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)
```

1. **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:

```
T1=c(1780, 1760, 1690, 1750, 1840, 1920, 1100, 1810, 1050, 1950)
t1m=mean(T1)
t1md=median(T1)
t1m
t1md
```

Output:

```
> t1m
[1]  1665

> t1md
[1] 1770
```

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:

```
T2=c(105.26,101.05,113.60,105.97,95.05,93.58,76.21,67.42,77.88,77.97,104.44,174.11)
t2m=mean(T2)
t2md=median(T2)
t2m
t2md
```

Output:

```
> t2m

[1]  99.37833


>t2md

[1]  98.05
```

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.

R-Code:

```
T3=c(1945.63, 1864.69, 1093.11, 1297.27, 918.15)
t3m=mean(T3)
t3md=median(T3)
t3m
t3md
```

Output:

```
>t3m
[1]  1423.77

>t3md
[1]  1297.27
```

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

R – Code:

```
d=read.table(header=TRUE,text="x              f
  1             5
          2             9
 3            12
  4            17
    5            14
  6            10
 7            6")
t4= rep(d$x, d$f)
multi.fun = function(fr)
{
  c(mean=mean(fr),sd=sd(fr))
}
multi.fun(t4)
```

Output:

```
     mean              sd
4.095890       1.668036
```

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.

R – Code:

```
d=read.table(header=TRUE,text="Miles          Villagers
 2                38
  4                104
    6                140
    8                78
        10                48
      13                42
     14                28
      16                24
     18                16
    20                2")
t5= rep(d$Miles, d$Villagers)
multi.fun = function(fr)
{
  c(mean=mean(fr),median=median(fr))
}
multi.fun(t5)
```

Output:

```
mean       median
7.857692   6.000000
```

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

R – Code:

```
d=read.table(header=TRUE,text="Size          Frequency
   6              3
      7              6
    8              9
    9              13
   10             8
 11             5
   12             4")
t6= rep(d$Size, d$Frequency)
multi.fun = function(fr)
{
  c(mean=mean(fr),sd=sd(fr))
}
multi.fun(t6)
```

Output:

```
mean             sd
9.000000   1.624284
```

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

R – Code:

```
mode=function(x)
{
ux= unique(x)
ux[which.max(tabulate(match(x,ux)))]
}
T7=c(14.8,14.2,13.8,13.5,14.0,14.2,14.3,14.6,13.9,14.0,14.1,13.2,13.0,14.2,13.5,13.0,12.8,13.9,1
4.8,15.0,12.8,13.4,13.2,14.0,13.8,13.9,14.0,14.0,13.9,14.8)
c(mean=mean(T7),median=median(T7),mode=mode(T7))
```

Output:

```
    mean      median          mode
13.88667   13.95000    14.00000
```

KUMARAGURU COLLEGE OF TECHNOLOGY

LABORATORY MANUAL

Experiment Number: 7

Lab Code : **U18MAI4201**

Lab : **Probability and Statistics**

Course / Branch : **B.E-CSE,ISE, B.Tech-IT**

Title of the Experiment : **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 ≥ 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 $O_i(i = 1,2,\ldots.n)$ be a set of observed frequencies and $E_i(i = 1,2,\ldots.n)$ 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 $\boldsymbol{\chi^2}$-value with the tabulated $\boldsymbol{\chi^2}$-value (with n – 1 d.f) and form the conclustion.

$\boldsymbol{\chi^2}$- test of Independence of Attributes

$\boldsymbol{\chi^2}$- 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 rsclasses. 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:

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

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

Test statistic is $\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$ 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

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

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: The digits occur equally frequently in the directory.

Alternative Hypothesis: The digits does not occur equally frequently in the directory

Level of significance: 5%

R-code:

```
frequ=c(1026,1107,997,996,1075,933,1107,972,964,853)
p=c(1/10,1/10,1/10,1/10,1/10,1/10,1/10,1/10,1/10,1/10)
a=chisq.test(frequ,p=c(1/10,1/10,1/10,1/10,1/10,1/10,1/10,1/10,1/10,1/10))
a
```

Output:

```
        Chi-squared test for given probabilities

data:  frequ
X-squared = 58.542, df = 9, p-value = 2.558e-09
```

Table value of $\chi^2_{0.05}$ for 9 d.f = 16.919

Conclusion:

$\chi^2 > \chi^2_{0.05}$,so we reject H_0 and conclude that the digits does not occur equally frequently in the directory

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.

Null Hypothesis: The arrival of trucks is equally distributed at the 0.05 level of significance
Alternative Hypothesis: The arrival of trucks is not equally distributed at the 0.05 level of significance

Level of significance: 5%

R-code:

```
frequ=c(52,51,56,47,60,57,59,61,57)

p=c(1/9,1/9,1/9,1/9,1/9,1/9,1/9,1/9,1/9)

a=chisq.test(frequ,p=c(1/9,1/9,1/9,1/9,1/9,1/9,1/9,1/9,1/9))

a
```

Output:

```
        Chi-squared test for given probabilities


data:  frequ

X-squared = 3.1, df = 8, p-value = 0.9279
```

Table value of $\chi^2_{0.05}$ for 8 d.f = 15.507

Conclusion:

$\chi^2 < \chi^2_{0.05}$,so we accept H_0 and conclude that 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: The nature of the area is not related to voting preference in the election

Alternative Hypothesis: H_1: the nature of the area is related to voting preference in the election

Level of significance: α = 5%

R-code:

```
rural = c(620,380)

urban= c(550,450)

office = as.data.frame(rbind(rural,urban))

chisq.test(office,simulate.p.value=T)
```

Output:

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

data: office

X-squared = 10.092, df = NA, p-value = 0.002499

Table value: χ^2= 3.841

Conclusion:

$\chi^2 > \chi_\alpha{}^2$, we reject H_0 hence we conclude that the nature of the area is related to voting preference in the election

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: The Drug is not effective.

Alternative Hypothesis: H_1: The Drug is effective

Level of significance: α = 5%

R-code:

```
cured = c(65,55)

notcured= c(35,45)

Drug = as.data.frame(rbind(cured,notcured))

chisq.test(Drug,simulate.p.value=T)
```

Output:

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

data: Drug

X-squared = 2.0833, df = NA, p-value = 0.1839

Table value: χ^2= 3.841

Conclusion:

$\chi^2<\chi_\alpha{}^2$, we accept H_0 hence we conclude that the drug is not effective.

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: There is no relation between smoking and literacy

Alternative Hypothesis: H_1: There is a relation between smoking and literacy

Level of significance: α = 5%

R-code:

```
literates = c(83,57)
illiterates= c(45,68)
smoke = as.data.frame(rbind(literates,illiterates))
chisq.test(smoke,simulate.p.value=T)
```

Output:

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

data: smoke

X-squared = 9.4757, df = NA, p-value = 0.001999

Table value: χ^2= 3.841

Conclusion:

$\chi^2 > \chi_\alpha{}^2$, we reject H_0 hence we conclude that there is a relation between smoking and literacy

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: There is no association between intelligence and economic conditions

Alternative Hypothesis: H_1: There is association between intelligence and economic conditions

Level of significance: α = 5%

R-code:

```
good = c(48,200,150,80)

notgood= c(52,180,190,100)

int = as.data.frame(rbind(good,notgood))

chisq.test(int,simulate.p.value=T)
```

Output:

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

data: intl

X-squared = 6.2168, df = NA, p-value = 0.1039

Table value: 3 df at 5% χ^2= 7.815

Conclusion:

$\chi^2 < \chi_\alpha{}^2$, we accept H_0 hence we conclude that there is no association between intelligence and economic conditions

STEP 3: PRACTICE/TESTING

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

   The Chi Square statistic is commonly used for testing relationships between categorical variables. The Chi-Square statistic is most commonly used to evaluate Tests of Independence when using a bivariate table.

2. **State the 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 ≥ 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$4 \leq k \leq 16$

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

   The Chi-square goodness of fit test is a statistical hypothesis test used to determine whether a variable is likely to come from a specified distribution or not. It is often used to evaluate whether sample data is representative of the full population.

4. **When do we use χ^2- test of Independence of Attributes?**
   $\chi 2$- test is used 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. The observations are arranged in the form of a matrix, called contingency table.

KUMARAGURU COLLEGE OF TECHNOLOGY
LABORATORY MANUAL

Experiment Number: 8

Lab Code : **U18MAI4201**

Lab : **Probability and Statistics**

Course / Branch : **B.E-CSE,ISE, B.Tech-IT**

Title of the Experiment : 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 th
e 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**

Null Hypothesis: H_0: There is no significant difference between machines.

Alternative Hypothesis: H_1: There is significant difference between machines.

R-code:

```
pieces=c(17,16,14,13,15,12,19,18,20,8,11,17)

machine=c(rep("A",4),rep("B",4),rep("C",4))

data=data.frame(pieces,machine)

data

results=aov(pieces~machine,data=data)
```

summary(results)

Output:

	pieces	machine
1	17	A
2	16	A
3	14	A
4	13	A
5	15	B
6	12	B
7	19	B
8	18	B
9	20	C
10	8	C
11	11	C
12	17	C

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
machine	2	8	4.00	3.61	0.764
Residuals	9	130	14.44		

Conclusion:

$F_\alpha = 19.38$, $F < F_\alpha$, so we accept the null hypothesis and conclude that there is no significant difference between machines.

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?

Null Hypothesis: H_0: There is no significant difference between machines.

Alternative Hypothesis: H_1: There is significant difference between machines.

R-code:

```
cotton=c(8,9,11,12,6,8,10,4,14,12,18,9,20,22,25,23)

machine=c(rep("A",4),rep("B",4),rep("C",4),rep("D",4))

data=data.frame(cotton,machine)

data

results=aov(cotton~machine,data=data)

summary(results)
```

Output:

```
   cotton machine
1       8       A
2       9       A
3      11       A
4      12       A
5       6       B
6       8       B
```

7 10 B

8 4 B

9 14 C

10 12 C

11 18 C

12 9 C

13 20 D

14 22 D

15 25 D

16 23 D

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
machine	3	540.7	180.23	25.22	1.81e-05 ***
Residuals	12	85.7	7.15		

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

Conclusion:

$F_\alpha = 3.49$, $F > F_\alpha$, so we reject the null hypothesis and conclude that there is significant difference between 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

Null Hypothesis: H_0: There is no significant difference between variety yields.

Alternative Hypothesis: H_1: There is significant difference between variety yields.

R-code:

```
variety=c(7,7,14,11,9,6,9,8,12,9,8,9,13,10,9,7,13,13,11,11,7,6,16,11,12,6,12,11,11,11)
plots=c(rep("I",10),rep("II",10),rep("III",10))
data=data.frame(variety,plots)
data
results=aov(variety~plots,data=data)
summary(results)
```

Output:

```
   variety plots
1        7     I
2        7     I
3       14     I
4       11     I
5        9     I
6        6     I
7        9     I
8        8     I
9       12     I
10       9     I
11       8    II
12       9    II
13      13    II
14      10    II
```

15 9 II

16 7 II

17 13 II

18 13 II

19 11 II

20 11 II

21 7 III

22 6 III

23 16 III

24 11 III

25 12 III

26 6 III

27 12 III

28 11 III

29 11 III

30 11 III

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
plots	2	8.87	4.433	1.555	0.533
Residuals	27	186.10	6.893		

Conclusion:

$F_\alpha = 19.46$, $F < F_\alpha$, so we accept the null hypothesis and conclude that there is no significant difference 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**

Null Hypothesis: H_0: There is no significant difference between fabric strength.

Alternative Hypothesis: H_1: There is significant difference between fabric strength

R-code:

```
strength=c(8.67,8.68,8.66,8.65,7.68,7.58,8.67,8.65,8.62,8.69,8.67,8.92,7.7,7.7,7.90,8.65,8.20,
8.60)

dye=c(rep("A",4),rep("B",5),rep("C",4),rep("D",5))

data=data.frame(strength,dye)

data

results=aov(strength~dye,data=data)

summary(results)
```

Output:

```
  strength dye
1     8.67   A
2     8.68   A
3     8.66   A
4     8.65   A
5     7.68   B
```

6	7.58	B
7	8.67	B
8	8.65	B
9	8.62	B
10	8.69	C
11	8.67	C
12	8.92	C
13	7.70	C
14	7.70	D
15	7.90	D
16	8.65	D
17	8.20	D
18	8.60	D

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
dye	3	0.6202	0.2067	1.023	0.412
Residuals	14	2.8304	0.2022		

Conclusion:

$F_\alpha = 3.34$, $F < F_\alpha$, so we accept the null hypothesis and conclude that there is no significant difference between fabric strength.

STEP 3: PRACTICE/TESTING

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

   1. Replication – Repetition of treatments under investigation
   2. Randomisation – Assigning treatments randomly to the experimental units
   3. Local control – Making the experimental units homogeneous and reducing the experimental error

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

1. Planning of the experiment
2. Obtaining relevant information from the experiment regarding the statistical hypothesis under study.
3. Making a statistical analysis of the data

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

The term Completely Randomized Design refers to the fact that a single variable factor of interest is controlled and its effect on the other elementary units is observed. In other words, the data are classified according to only one criterion in one-way classification.

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

Analysis of variance, or ANOVA, is a statistical method that separates observed variance data into different components to use for additional tests. A one-way ANOVA is used for three or more groups of data, to gain information about the relationship between the dependent and independent variables.

KUMARAGURU COLLEGE OF TECHNOLOGY

LABORATORY MANUAL

Experiment Number: 9

Lab Code : **U18MAI4201**

Lab : **Probability and Statistics**

Course / Branch : **B.E-CSE,ISE, B.Tech-IT**

Title of the Experiment : 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, $F_{0.054,12=3.26}$

$$F_{0.053,12=3.49}$$

$F1 = 6.54 > F_{0.054,12=3.26}$, hence we reject H_{01} and conclude that the 5 workers differ with respect to mean productivity.

$F2 = 18.388 > F_{0.053,12=3.49}$, 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.

H_{01} : There is no difference between the sales in 3 seasons

H_{02} : There is no difference between the sales of the 4 salesman

R-code:

```
a=c(45,43,39,40,41,39,38,45,41,37,38,41)
f=c("Sum","Win","Mon")
k=3
r=4
season=gl(k,1,r*k,factor(f))
season
salesman=gl(r,k,k*r)
salesman
av = aov(a ~ season+salesman)
summary(av)
```

Output:

```
>season
[1] Sum Win Mon Sum Win Mon Sum Win Mon Sum Win Mon
Levels: Sum Win Mon
>salesman
[1] 1 1 1 2 2 2 3 3 3 4 4 4
Levels: 1 2 3 4
>summary(av)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
season	2	8.17	4.083	1.8709	0.611

salesman	3	22.92	7.639	1.000	0.455
Residuals	6	45.83	7.639		

Conclusion:

From F-table, $F0.056,2=19.3$; $F0.053,6=4.76$

$F1$ = 1.8709 <$F0.056,2=19.3$, hence we accept H_{01} and conclude that there is no difference between the sales in 3 seasons

$F2$= 1.000 <$F0.053,6=4.76$, hence we accept H_{02} and conclude that there is no difference between the sales of the 4 salesman

Task 2

Four different, though supposedly equivalent, forms of a standardized reading achievementtest 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.

H_{01} : There is no difference between the forms.

H_{02} : There is no difference between the performances of the students.

R-code:

```
a=c(75,83,86,73,73,72,61,67,59,56,53,62,69,70,72,79,84,92,88,95)
f=c("A","B","C","D")
k=4
r=5
form=gl(k,1,r*k,factor(f))
form
student=gl(r,k,k*r)
student
av = aov(a ~ form+student)
summary(av)
```

Output:

```
>form
[1] A B C D A B C D A B C D A B C D A B C D
Levels: A B C D
>student
[1] 1 1 1 1 2 2 2 2 3 3 3 3 4 4 4 4 5 5 5 5
Levels: 1 2 3 4 5
>summary(av)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
form	3	43.0	14.3	1.9790	0.685
student	4	2326.7	581.7	20.572	2.65e-05 ***
Residuals	12	339.3	28.3		

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

Conclusion:

From F-table, $F0.0112,3 = 27.05$; $F0.014,12 = 5.41$

$F1 = 1.979 < F0.0112,3 = 27.05$, hence we accept H_{01} and conclude that there is no difference between the forms.

$F2 = 20.572 > F0.014,12 = 5.41$, hence we reject H_{02} and conclude that there is a significant difference between the performances of the students.

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.

H_{01} : There is no difference between the performance of the detergents.

H_{02} : There is no difference between the engines.

R-code:

```
a=c(45,47,48,42,43,46,50,37,51,52,55,49)
f=c("A","B","C","D")
k=4
r=3
detergent=gl(k,1,r*k,factor(f))
detergent
engine=gl(r,k,k*r)
engine
av = aov(a ~ detergent+engine)
summary(av)
```

Output:

```
>detergent
        [1] A B C D A B C D A B C D
        Levels: A B C D
>engine
[1] 1 1 1 1 2 2 2 2 3 3 3 3
Levels: 1 2 3
>summary(av)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
detergent	3	110.92	36.97	11.78	0.00631 **
engine	2	135.17	67.58	21.53	0.00183 **

Residuals 6 18.83 3.14

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

Conclusion:

From F-table, $F0.01(3,6) = 9.78$; $F0.012,6 = 10.92$

$F1 = 11.78 > F0.013,6 = 9.78$, hence we reject H_{01} and conclude that there is a difference between the performance of the detergents.

$F2 = 21.53 > F0.012,6 = 10.92$, hence we reject H_{02} and conclude that there is a difference between 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.

H_{01} : There is no significant difference between the observers

H_{02} : There is no significant difference between the consignments

R-code:

```
a=c(9,12,11,12,10,11,10,13,9,9,10,11,10,11,12,14,11,10,11,12,11,10,10,10)
f=c("1","2","3","4")
k=4
r=6
observer=gl(k,1,r*k,factor(f))
observer
consignment=gl(r,k,k*r)
consignment
av = aov(a ~ observer+consignment)
summary(av)
```

Output:

>observer

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

Levels: 1 2 3 4

>consignment

[1] 1 1 1 1 2 2 2 2 3 3 3 3 4 4 4 4 5 5 5 5 6 6 6 6

Levels: 1 2 3 4 5 6

>summary(av)

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
observer	3	13.125	4.375	5.000	0.0134 *
consignment	5	9.708	1.942	2.219	0.1064
Residuals	15	13.125	0.875		

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

Conclusion:

From F-table, $F_{0.053,15} = 3.29$; $F_{0.055,15} = 2.90$

$F1 = 5.0000 > F_{0.053,15} = 3.29$, hence we reject H_{01} and conclude that there is a significant difference between the observers.

$F2 = 2.219 < F_{0.055,15} = 2.90$, hence we accept H_{02} and conclude that there is no significant difference between the consignments

STEP 3: PRACTICE/TESTING

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

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 – i.e., $N=kr$ entries.

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

CRD – One Way Classification ; RBD – Two Way Classification

- RBD is more efficient/accurate than CRD for most types of experimental work.
- RBD is more flexible than CRD since no restrictions are placed on the number of treatments or the number of replications.

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

- This design is more efficient/accurate than CRD, i.e., it has less experimental error.
- This design is more flexible. ie. no restrictions are placed on the number of treatments or the number of replications

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

An ANOVA test is a way to find out if survey or experiment results are significant. They help us to decide whether we should accept or reject the null hypothesis. This technique is used to compare means and the relative variance between them. It is used when three or more populations or samples are to be compared.

KUMARAGURU COLLEGE OF TECHNOLOGY

LABORATORY MANUAL

Experiment Number: 10

Lab Code **: U18MAI4201**

Lab : Probability and Statistics

Course / Branch : B.E-CSE,ISE, B.Tech-IT

Title of the Experiment: Control charts for variables (mean and range chart)

STEP 1: INTRODUCTION

OBJECTIVES OF THE EXPERIMENT

To plot $\bar{X}$- 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 ($\bar{X}$)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. $\bar{X}$chart LCL:$\bar{X}-A2\bar{R}$ UCL: $\bar{X}+A2\bar{R}$

2. R-Chart LCL: $D3\bar{R}$ UCL: $D4\bar{R}$

Procedure to plot $\bar{X}$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 *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 $\bar{X}$ 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 $\bar{X}$ chart:**

```
Xbarchart= qcc(data = A,

type = "xbar",

sizes = 5,

title = "X-bar Chart ",

plot = TRUE)
```

Output:

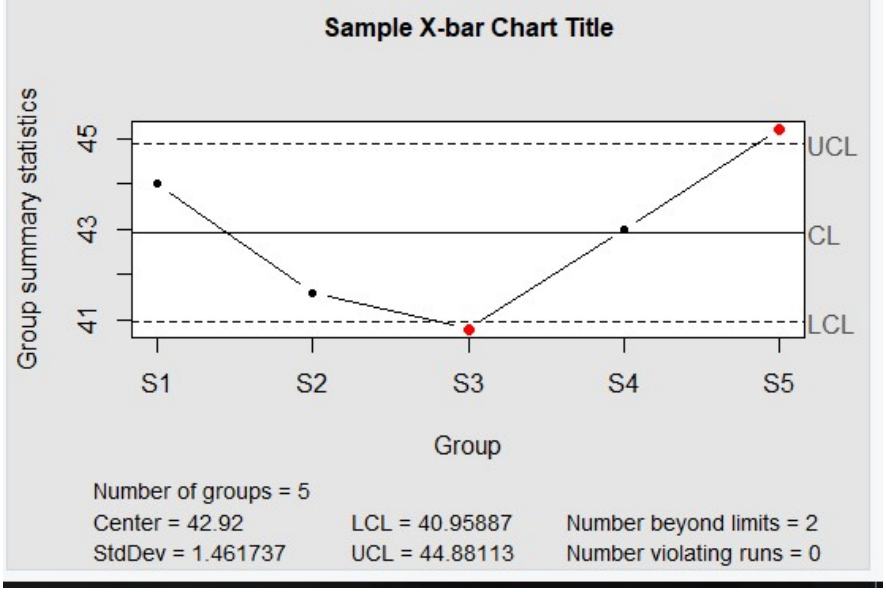
```
  V1 V2 V3 V4 V5

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

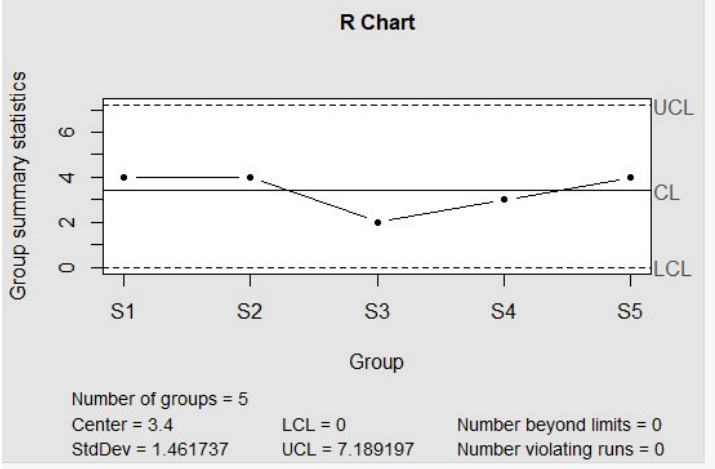


For R chart:

R-code:

```
rchart = qcc(data = A,
type = "R",
sizes = 5,
title = "R Chart",
plot = TRUE)
```

Output:



Conclusion:

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

R-Code:

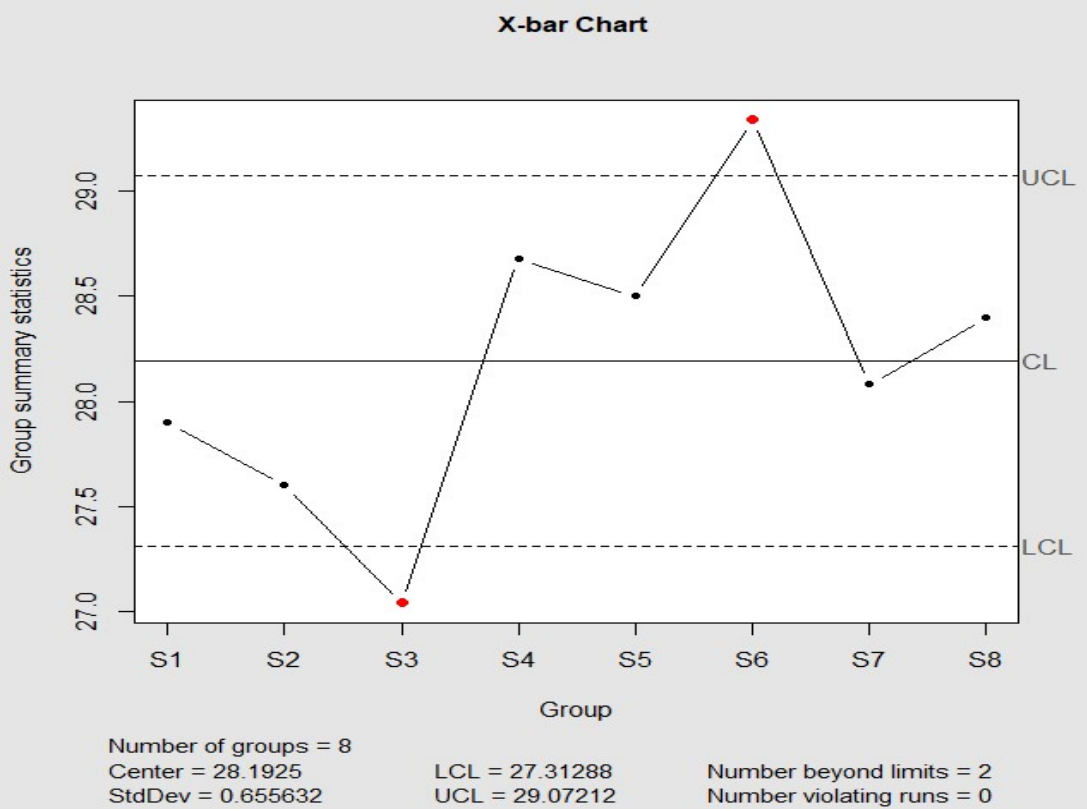
```
S1=c(27.5,28.5,28,26.9,28.6)
S2=c(27.4,26.9,26,28.7,29.0)
S3=c(25.4,26.9,28,26.7,28.2)
S4=c(28.5,28,29.2,29,28.7)
S5=c(28.5,29,28.5,28.5,28)
S6=c(28.9,29.5,30,29.4,28.9)
S7=c(28,28.5,27.8,28,28.1)
S8=c(28.4,28.5,28.4,28,28.7)
A= as.data.frame(rbind(S1,S2,S3,S4,S5,S6,S7,S8))
A
```

Output:

```
> A
    V1   V2   V3   V4   V5
S1 27.5 28.5 28.0 26.9 28.6
S2 27.4 26.9 26.0 28.7 29.0
S3 25.4 26.9 28.0 26.7 28.2
S4 28.5 28.0 29.2 29.0 28.7
S5 28.5 29.0 28.5 28.5 28.0
S6 28.9 29.5 30.0 29.4 28.9
S7 28.0 28.5 27.8 28.0 28.1
S8 28.4 28.5 28.4 28.0 28.7
```

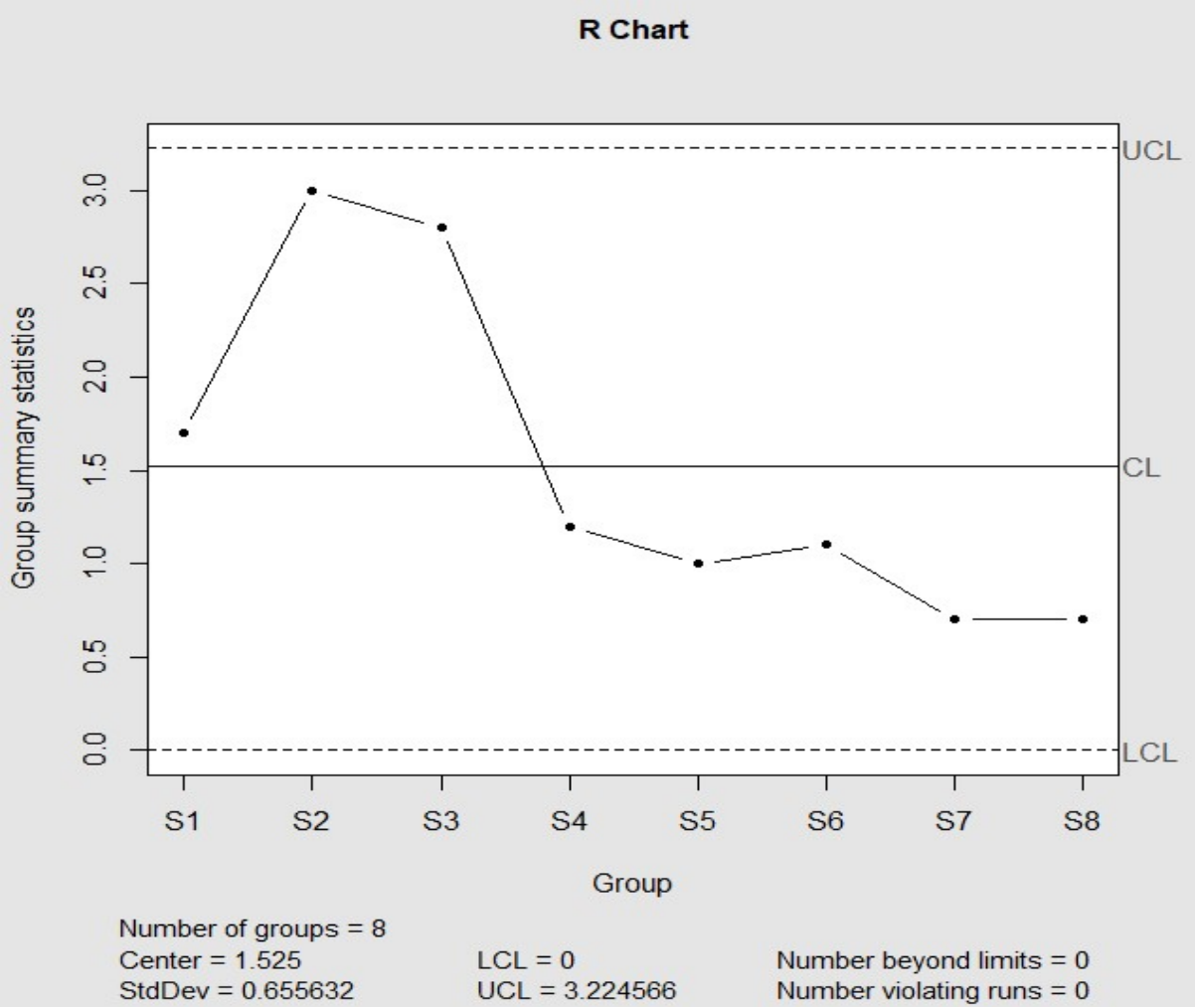
$\bar{X}$ chart:

Xbarchart= qcc(data = A,type = "xbar",sizes = 5,title = "X-bar Chart ",plot = TRUE)



R – Chart:

rchart = qcc(data = A,type = "R",sizes = 5,title = "R Chart",plot = TRUE)



Conclusion:

In $\bar{X}$ 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 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**

R-Code:

```
S1=c(47,49,50,44,45)

S2=c(52,55,47,56,50)

S3=c(48,53,51,50,53)

S4=c(49,49,49,53,45)

S5=c(50,53,48,52,47)

S6=c(55,55,50,53,57)

S7=c(50,51,53,46,50)

S8=c(54,54,52,54,56)

S9=c(49,55,54,49,53)

S10=c(53,50,54,47,51)

A= as.data.frame(rbind(S1,S2,S3,S4,S5,S6,S7,S8,S9,S10))

A
```

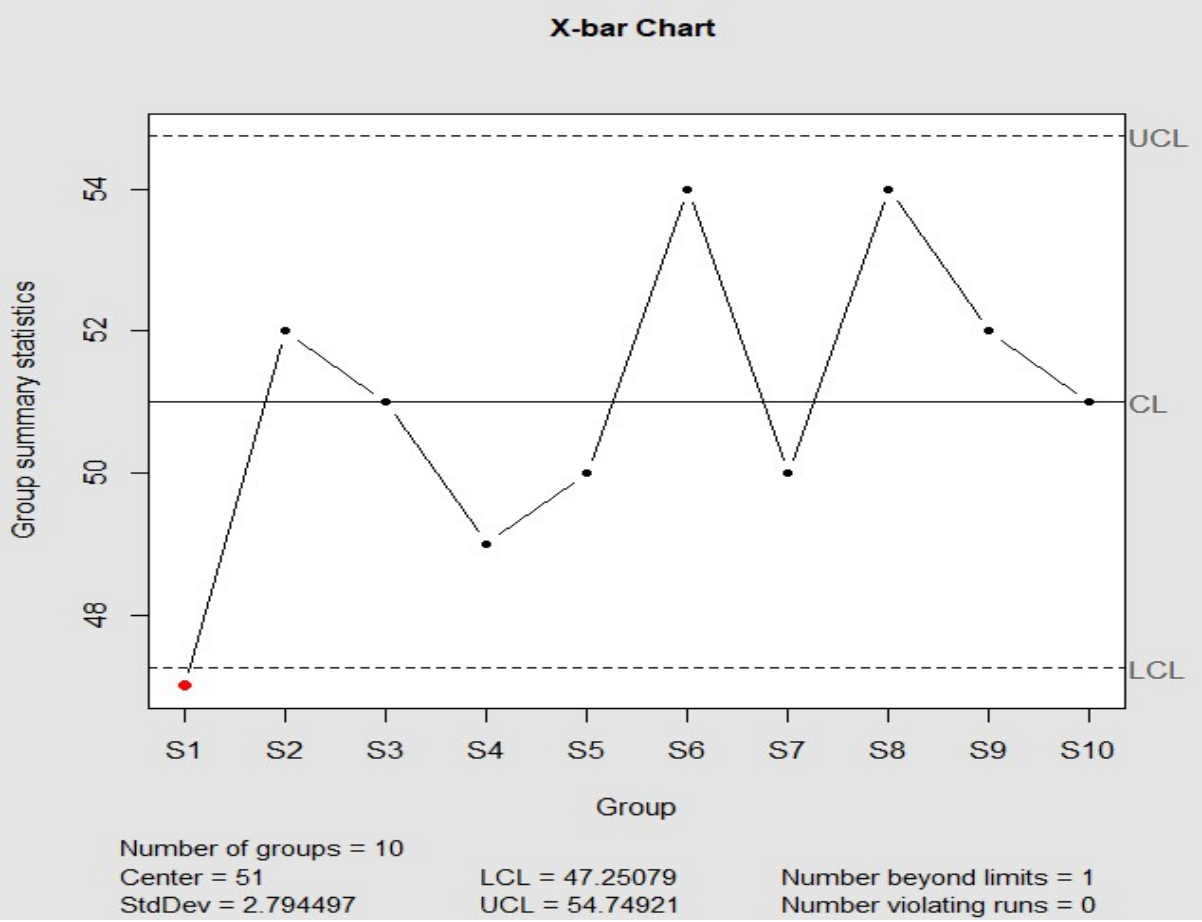
Output:

```
> A
```

	V1	V2	V3	V4	V5
S1	47	49	50	44	45
S2	52	55	47	56	50
S3	48	53	51	50	53
S4	49	49	49	53	45
S5	50	53	48	52	47
S6	55	55	50	53	57
S7	50	51	53	46	50
S8	54	54	52	54	56
S9	49	55	54	49	53
S10	53	50	54	47	51

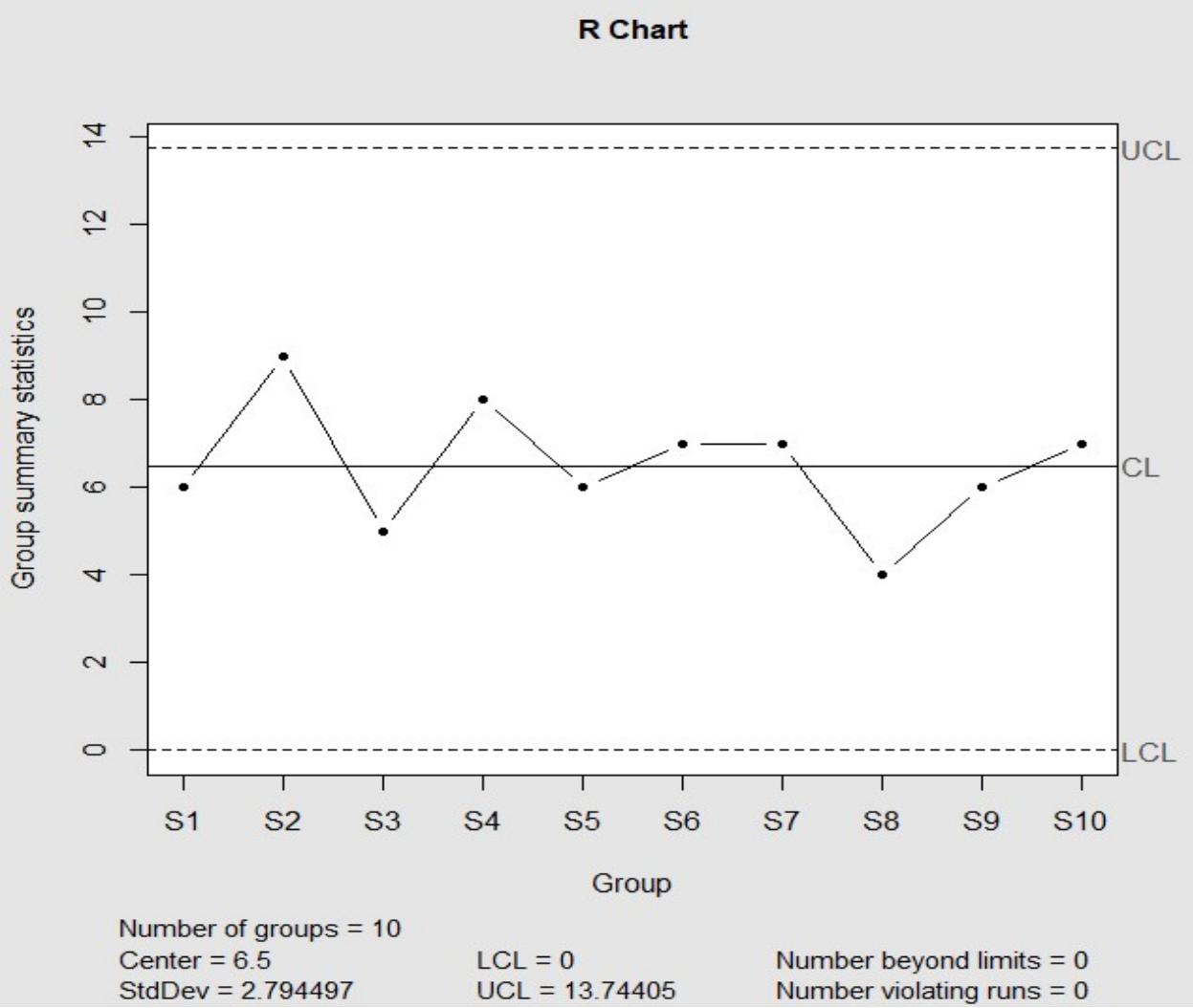
$\bar{X}$ chart:

```
Xbarchart= qcc(data = A,type = "xbar",sizes = 5,title = "X-bar Chart ",plot = TRUE)
```



R – Chart:

rchart = qcc(data = A,type = "R",sizes = 5,title = "R Chart",plot = TRUE)



Conclusion:

In $\bar{X}$ chart, one point lies 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 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

R-Code:

```
S1=c(469.92,468.67,479.76,454.38,469.58,454.46)

S2=c(457.34,454.37,475.28,453.46,480.03,480.4)

S3=c(473.96,459.26,460.42,462.04,450.6,451.52)

S4=c(480.06,469.86,456.42,460.63,465.66,466.99)

S5=c(467.46,476.56,474.01,465.34,475.27,462.97)

S6=c(473.06,475.86,472.97,454.93,470.73,466.24)

S7=c(456.27,476.37,479.50,459.86,470.73,452.35)

A= as.data.frame(rbind(S1,S2,S3,S4,S5,S6,S7))

A
```

Output:

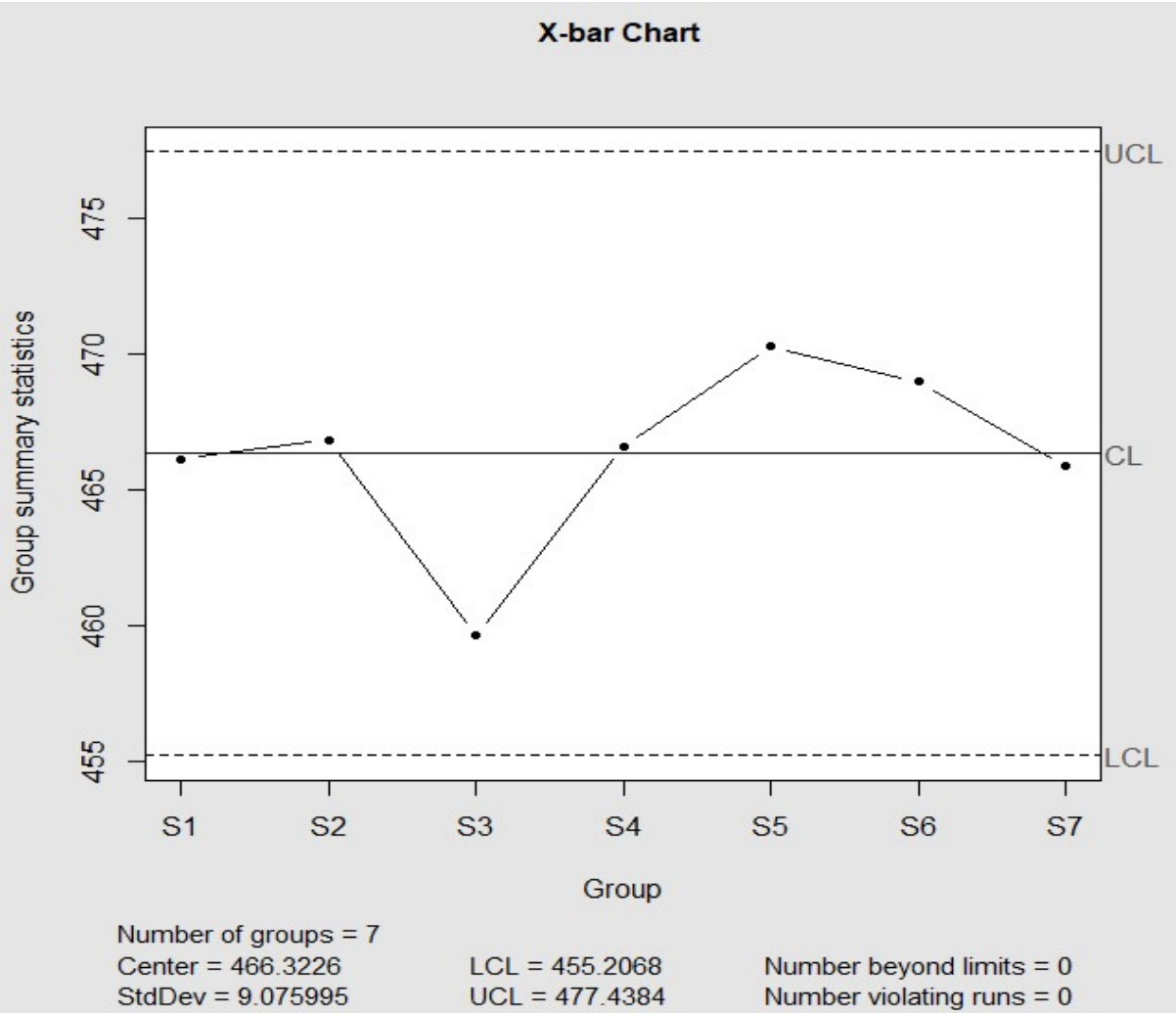
```
> A
      V1     V2     V3     V4     V5     V6
S1 469.92 468.67 479.76 454.38 469.58 454.46
S2 457.34 454.37 475.28 453.46 480.03 480.40
S3 473.96 459.26 460.42 462.04 450.60 451.52
S4 480.06 469.86 456.42 460.63 465.66 466.99
S5 467.46 476.56 474.01 465.34 475.27 462.97
```

S6 473.06 475.86 472.97 454.93 470.73 466.24

S7 456.27 476.37 479.50 459.86 470.73 452.35

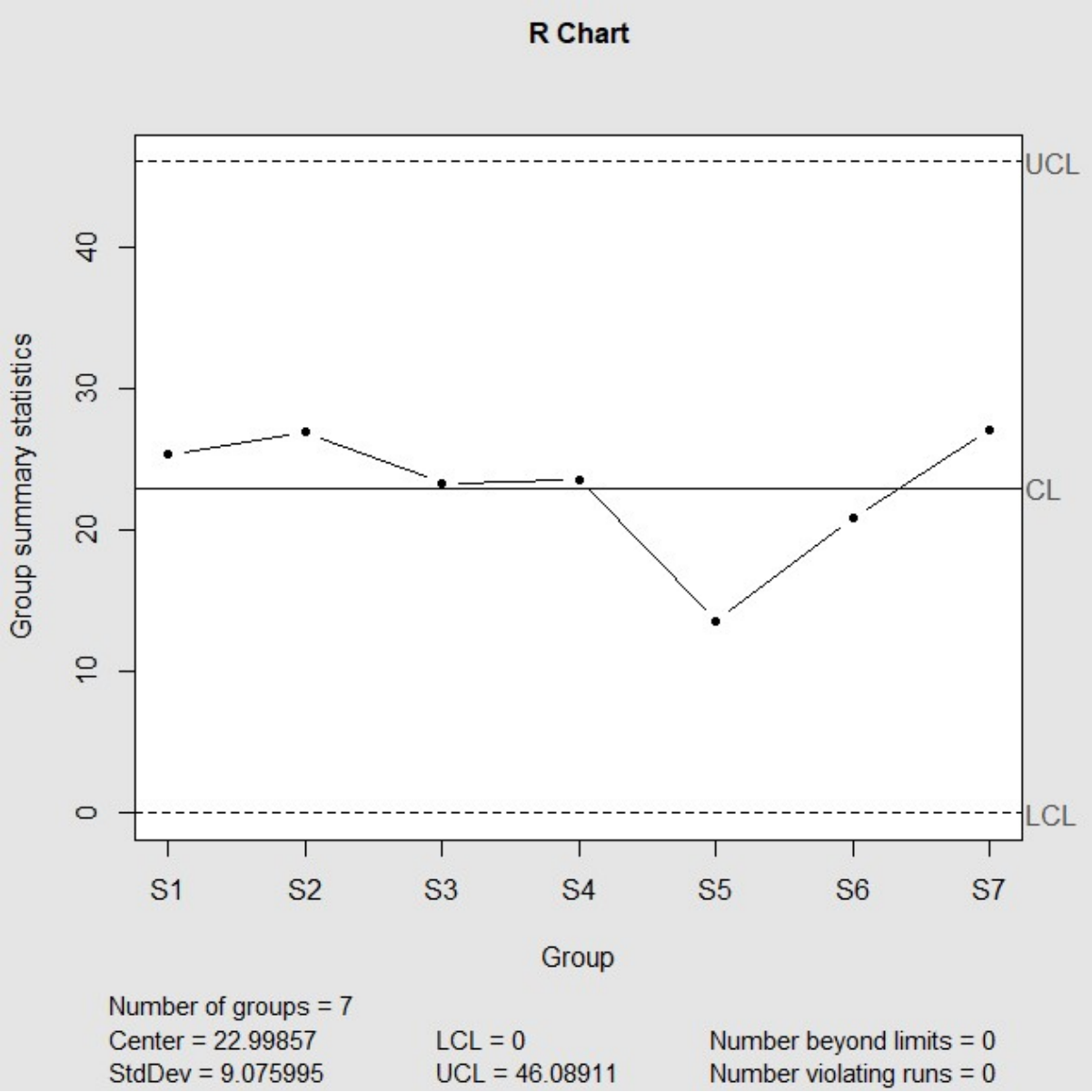
$\bar{X}$ chart:

Xbarchart= qcc(data = A,type = "xbar",sizes = 5,title = "X-bar Chart ",plot = TRUE)



R – Chart:

rchart = qcc(data = A,type = "R",sizes = 5,title = "R Chart",plot = TRUE)



Conclusion:

In $\bar{X}$ chart, all points are within the control limits, so as far as sample mean is concerned, the system is under 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 under control

STEP 3: PRACTICE/TESTING

1. **Define Statistical Quality Control.**

- The technique of applying statistical methods based on sampling to establish quality standards and to maintain it in the most economical manner.
- The objective of the SQC is to devise the fact that even after the quality standards have been specified, some variation in quality e statistical methods that isolate assignable variation from random variation and enable us to detect, identify and eliminate the assignable causes of variation.

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

Control Chart is a important statistical tool used for the study and control of repetitive processes. A control chart accepts the normal variation due to chance causes but eliminates entirely the errors due to assignable causes.

Two types of Control Charts:

1. Contorl Charts for Variables - $\bar{X}$ Chart, R Chart
2. Control Charts for Attributes - p Chart, np Chart, c Chart

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

Mean Chart:

Lower Control Limit**:** $\bar{\bar{X}} - A2R -$

Control Limit: $\bar{\bar{X}}$

Upper Control Limit: $\bar{\bar{X}} + A2R-$

R Chart:

Lower Control Limit**:** $D3R-$

Control Limit: $R-$

Upper Control Limit: $D4R-$

WORKSHEET 1

1. **What is a dataframe?**

   A data frame is a table or a two-dimensional array-like structure in which each column contains values of one variable and each row contains one set of values from each column.

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

The Characteristics of a data frame are

- The column names should be non-empty.
- The row names should be unique.
- The data stored in a data frame can be of numeric, factor or character type.
- Each column should contain same number of data items.

3. **How would you extract the subsets of all MQ students from the data frame in task 2?**

```
details = read.csv("Studentdetails1.csv")

MqStudent=subset(details,SEATCATG==”MQ”)

MqStudent
```

KUMARAGURU COLLEGE OF TECHNOLOGY

LABORATORY MANUAL

Experiment Number: 3

Lab Code : **U18MAI4201**

Lab : **Probability and Statistics**

Course / Branch : **B.E-CSE,ISE, B.Tech-IT**

Title of the Experiment : **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
6.	**To construct the regression plot of y on x** 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.

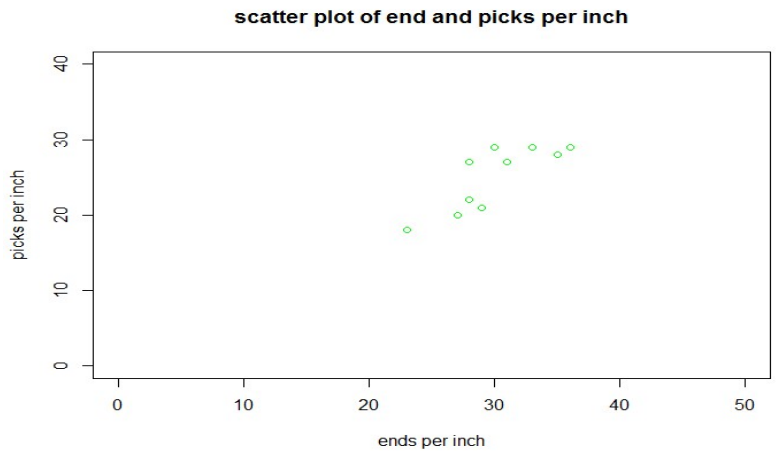
x	**23**	**27**	**28**	**28**	**29**	**30**	**31**	**33**	**35**	**36**
y	**18**	**20**	**22**	**27**	**21**	**29**	**27**	**29**	**28**	**29**

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 ="X -axis",ylab ="Y-
axis",xlim=c(0,50),ylim=c(0,40),col=c("green"),main="Introduction atoScatterPlot")
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
```

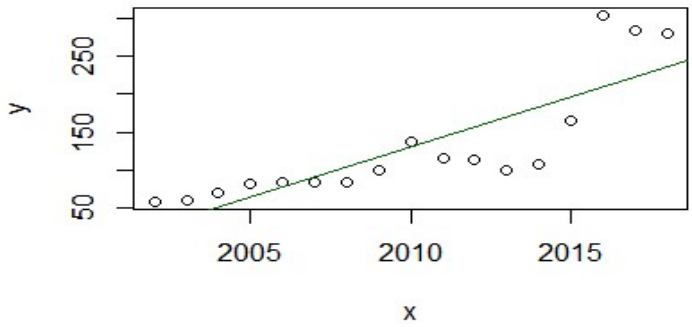
Regression plot of y on x

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.

Solution:

R-Code:

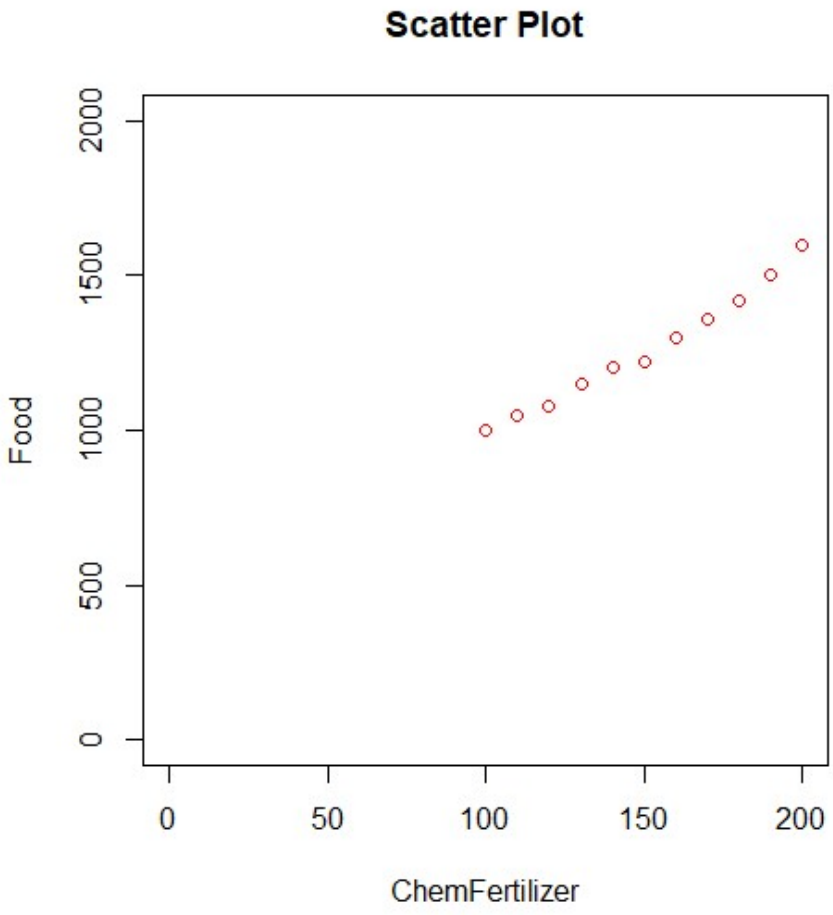
x=c(100,110,120,130,140,150,160,170,180,190,200,210,220,230)

```
y=c(1000,1050,1080,1150,1200,1220,1300,1360,1420,1500,1600,1650,1650,1650)
plot(x,y,xlab="ChemFertilizer",ylab="Food",xlim=c(0,200),ylim=c(0,2000),col=c("red"),mai
n="Scatter Plot")
r=cor(x,y)
r
```

Output:

```
>r
[1] 0.991053
```

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

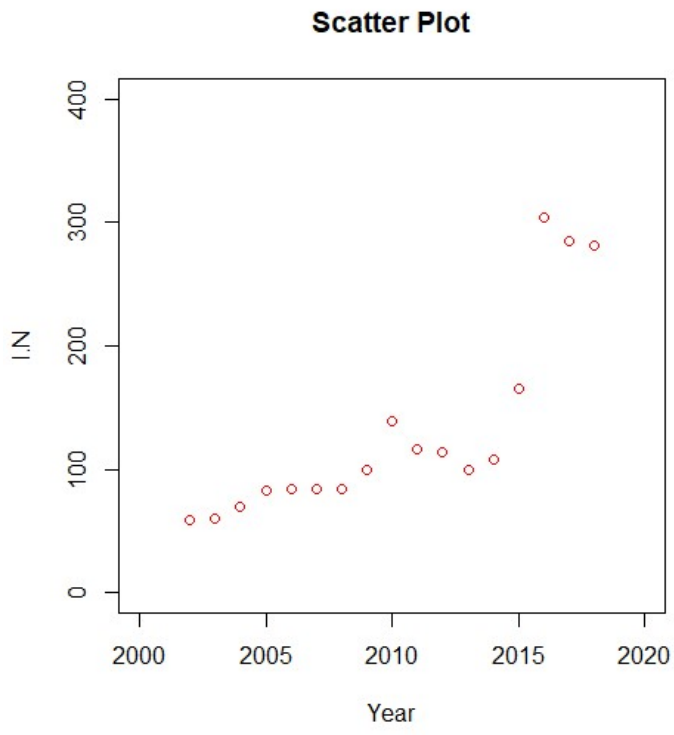
R-CODE:

```
x=c(2002,2003,2004,2005,2006,2007,2008,2009,2010,2011,2012,2013,2014,2015,2016,2017
,2018)

y=c(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)

plot(x,y,xlab="Year",ylab="I.N",xlim=c(2000,2020),ylim=c(0,400),col=c("red"),main="Scatt
er Plot")

r=cor(x,y)

r
```

Output:

```
>r

[1] 0.8306042
```

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.

Solution:

R-CODE:

```
x=c(14,15,1,6,5,3,10,2,4,9,7,8,12,13,11)
y=c(15,13,11,3,5,8,4,7,10,2,1,6,9,12,14)
z=c(12,11,6,4,9,8,1,2,3,10,5,7,15,14,13)
rankAB=cor(x,y,method="spearman")
rankBC=cor(y,z,method="spearman")
rankAC=cor(x,z,method="spearman")
rankAB
rankBC
rankAC
```

Output:

```
>rankAB
[1] 0.3857143
>rankBC
[1] 0.5357143
>rankAC
[1] 0.6571429
```

Conclusion:

Judges A and C have the nearest approach to common taste in food.

Task 4:

Import the excel file "Cotton prices-International and Domestic2" and the find the correlation between the monthly international average prices and the monthly domestic average prices

R-CODE:

```
data = read.csv("Cotton prices-International and Indian.csv")

print(data)
```

Output:

```
   Cotlook.A.Minimum Cotlook.A.Maximum Range Cotlook.A...Average
1              79.85             85.30  5.45                81.95
2              79.40             82.20  2.80                80.87
3              81.85             84.80  2.95                83.37
4              83.10             90.35  7.25                85.51
5              88.80             90.90  2.10                89.71
6              91.40             98.85  7.45                94.45
7              90.60             95.70  5.10                92.68
8              89.40             95.10  5.70                92.74
9              88.80             96.65  7.85                93.08
10             91.15             93.95  2.80                92.60
11             89.15             97.35  8.20                92.59
12             88.35             91.45  3.10                89.95
13             85.40             93.15  7.75                89.33
14             83.75             85.60  1.85                84.64
15             85.15             89.70  4.55                87.49
16             88.05             94.45  6.40                90.96
17             91.95             95.75  3.80                94.05
18             93.30             98.90  5.60                96.93
19             92.20             97.75  5.55                94.20
20             89.40             95.80  6.40                92.71
21             89.30             93.70  4.40                90.90
22             79.60             88.40  8.80                83.84
23             72.15             76.05  3.90                74.04
24             69.95             76.15  6.20                73.38
25             69.65             71.45  1.80                70.35
26             65.90             70.00  4.10                67.53
27             66.00             70.25  4.25                68.38
28             65.30             68.75  3.45                67.35
29             67.05             71.75  4.70                69.84
30             67.20             71.25  4.05                69.35
31             69.55             73.95  4.40                71.72
32             71.05             74.70  3.65                72.86
33             71.25             74.35  3.10                72.36
34             70.65             74.80  4.15                72.35
35             69.85             74.10  4.25                71.82
36             66.40             70.25  3.85                68.74
37             66.65             70.85  4.20                69.03
38             68.30             70.55  2.25                69.22
39             69.50             71.70  2.20                70.39
40             67.70             69.95  2.25                68.75
41             65.05             68.95  3.90                66.57
42             64.05             66.50  2.45                65.46
43             66.40             71.70  5.30                69.28
44             68.80             72.95  4.15                70.28
45             71.80             76.15  4.35                74.10
46             74.85             85.39 10.54                81.07
47             75.70             85.85 10.15                80.26
48             75.00             80.65  5.65                77.87
49             76.55             80.35  3.80                78.52
```

50	76.95	81.15	4.20	78.92
51	78.20	80.70	2.50	79.53
52	79.65	84.25	4.60	82.33
53	84.10	86.80	2.70	85.16
54	85.75	88.10	2.35	86.84
55	84.60	88.80	4.20	87.04
56	86.40	94.90	8.50	88.64
57	82.60	87.70	5.10	84.66
58	82.20	85.05	2.85	84.09
59	77.40	81.35	3.95	79.36
60	78.55	84.70	6.15	80.59
61	77.60	80.40	2.80	78.60
62	79.00	81.60	81.60	61.80

	Shankar.6.Minimum	Shankar.6.Maximum	Range.1	Shankar.6.Average
1	32900	34400	1500	33450
2	33000	33800	800	33564
3	33300	34200	900	33764
4	33600	34300	700	33771
5	33900	37200	3300	35013
6	37000	39300	2300	38275
7	36700	39400	2700	38139
8	37000	38600	1600	37742
9	38500	41500	3000	39892
10	41000	43200	2200	42370
11	42400	49000	6600	45968
12	46900	48900	2000	47805
13	41000	48500	7500	44776
14	38800	40900	2100	39935
15	38500	40400	1900	39284
16	40200	42800	2600	42015
17	41800	43200	1400	42565
18	41500	42400	900	41943
19	41400	42900	1500	42038
20	40700	43200	2500	42065
21	41200	42900	1700	42044
22	39500	42900	3400	41542
23	39000	40500	1500	39835
24	34700	39900	5200	38360
25	32700	34000	1300	33448
26	32400	33200	800	32812
27	32900	33300	400	33146
28	29800	32900	3100	31300
29	30100	31300	1200	30678
30	30700	32600	1900	31122
31	32200	34200	2000	33296
32	34200	35500	1300	34922
33	33200	35000	1800	34232
34	33800	34600	800	34293
35	33500	34700	1200	33992
36	33000	35500	2500	34672
37	31800	32900	1100	32472
38	32000	32500	500	32209
39	32400	34000	1600	33223
40	33400	34000	600	33672
41	33100	33800	700	33452
42	32100	33200	1100	32676
43	32800	34700	1900	33975
44	34700	36800	2100	35315
45	36700	42700	6000	39456
46	42700	48500	5800	45896
47	43900	47800	3900	46269
48	43000	48000	5000	45125
49	37700	44500	6800	41233
50	37700	40000	2300	38728
51	38600	39600	1000	39007
52	40000	42600	2600	41256
53	41900	43000	1100	42482
54	42600	43700	1100	43085
55	42100	44000	1900	42967

56	41600	43000	1400	42396
57	42300	43100	800	42642
58	41800	43300	1500	42362
59	42200	42600	400	42323
60	38700	42300	3600	40829
61	37800	39000	1200	38468
62	37200	38100	900	35861

#The correlation between the monthly international average prices and the monthly domestic average prices

Rcode:

```
interNatAvg = dataCot[, 5]

domestAvg = dataCot[, 9]

r=cor(interNatAvg, domestAvg)

r
```

Output:

```
>r

[1] 0.6859373
```

Task 5:
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:

```
x=c(54,85,86,50,42,75,65,56)

y=c(2.45,1.21,1.20,2.84,3.25,1.86,1.90,2.32)

regyx=lm(y~x)

regyx

regxy=lm(x~y)

regxy


#when x=60
```

```
y=4.91906-0.04351*60

y

#when y=1.5

x=111.03-22.03*1.50

x
```

OUTPUT:

```
>regyx
Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)   x
   4.91906        -0.04351
```

The Regression Line of y on x is -0.04351x + 4.91906

```
>regxy=lm(x~y)
>regxy
Call:
lm(formula = x ~ y)

Coefficients:
(Intercept)                  y
     111.03            -22.03
```

The Regression Line of x on y is -22.03y + 111.03

```
>y
[1] 2.30846
>x
[1] 77.985
```

Task 6
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:

```
x=c(91,97,108,121,67,124,51,73,111, 57)

y=c(71,75,69,97,70,91,39,61,80,47)

regyx=lm(y~x)

regxy=lm(x~y)

regyx

regxy


#when x=150

y=14.8113+0.6132*150

y

plot()

abline(lm(pur~sal),col="red")
```

OUTPUT:

```
>regyx
Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)                x
   14.8113           0.6132
```

The Regression Line of y on x is 0.6132x + 14.8113

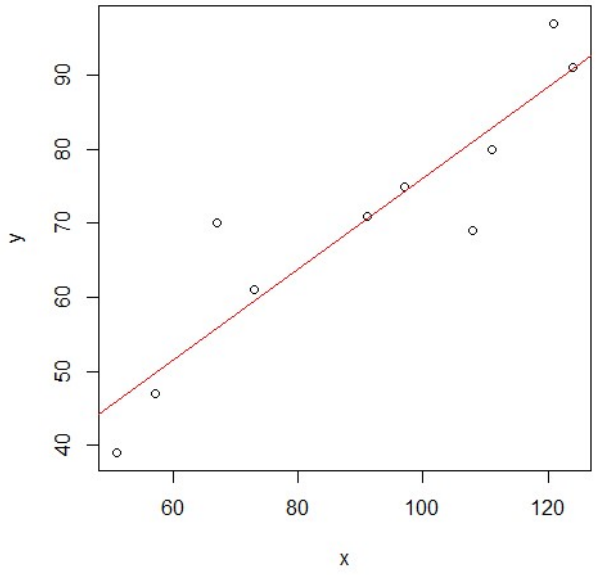
```
>regxy
Call:
lm(formula = x ~ y)

Coefficients:
(Intercept)                y
     -5.188            1.360
```

The Regression Line of x on y is 1.360y – 5.188

>y1

[1] 106.7913

Plot:



Task 7
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:

```
x=c(40,34,28,30,44,38,31)
y=c(32,39,26,30,38,34,28)
regyx=lm(y~x)
```

```
regyx
regxy=lm(x~y)
regxy

#When x=36
y=11.8703+0.5874*36
y
```

OUTPUT:

```
>regyx
Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)                x
   11.8703          0.5874
```

The Regression Line of y on x is 0.5874x + 11.8703

```
>regxy
Call:
lm(formula = x ~ y)

Coefficients:
(Intercept)                y
     7.6968          0.8419
```

The Regression Line of x on y is 0.8419y + 7.6968

```
> y

[1] 33.0167
```

Task 8

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:

```
x=c(56,42,36,47,49,42,60,72,63,55)
y=c(147,125,118,128,145,140,155,160,149,150)
regyx=lm(y~x)
regyx
regxy=lm(x~y)
regxy

#When x=45
y=83.76+1.11*45
y

#When y=172
x=-49.2958+0.7163*172
x
```

OUTPUT:

```
>regyx
Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)                x
      83.76             1.11
```

The Regression Line of y on x is 1.11x + 83.76

```
>regxy
Call:
lm(formula = x ~ y)

Coefficients:
(Intercept)                y
   -49.2958           0.7163
```

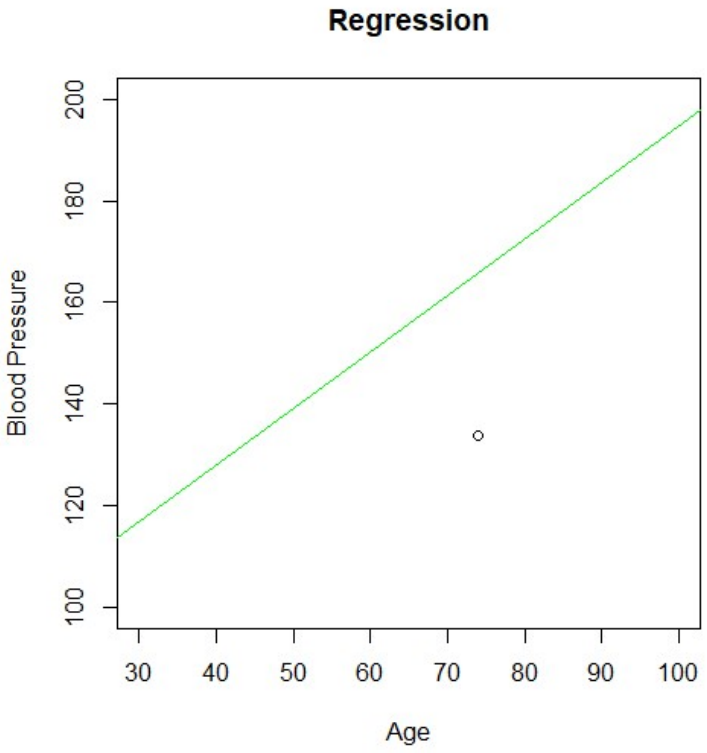
The Regression Line of x on y is 0.7163y – 49.2958

>y
[1] 133.71

> x
[1] 73.9078

```
plot(x,y,xlab="Age",ylab="Blood
Pressure",xlim=c(30,100),ylim=c(100,200),main="Regression")
abline(regyx,col="green")
```

Plot:



STEP 3: PRACTICE/TESTING

1. **Define correlation.**

   Correlation is a statistic that measures the degree to which two variables move in relation to each other.

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

   1. Scatter Diagram Method
   2. Karl Pearson's Correlation Coefficient
   3. Spearman's Rank Correlation Coefficient

3. **Explain scatter diagram.**

   The scatter diagram is a technique used to examine the relationship between both the axis (X and Y) with one variable. In the graph, if the variables are correlated, the point will drop along a curve or line. A scatter diagram or scatter plot, is used to give an idea idea of the nature of relationship.

4. **Define regression.**

Regression is the measure of the average relationship between two or more variables in terms of the original units of the data. It provides a mechanism for predicting or forecasting.

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

If two variables X and Y are correlated, we see that the scatter diagram will be more or less concentrated around a curve, called the curve of regression. If this curve is a straight line, then it is called line of regression.

The equation of the line of regression of Y on X is

$$y-\bar{y}=\text{byx}(x-\bar{x})$$

The equation of the line of regression of X on Y is

$$x-\bar{x}=\text{bxy}(y-\bar{y})$$

6. Mention some properties of regression lines.

Regression coefficients values remain the same. Since shifting of origin takes place because of the change of scale.

If there are two lines of regression. Both of these lines intersect at a specific point [x', y']. Variables x and y are taken into consideration. According to the property, the intersection of both the lines of regression i.e. y on x and y is [x', y']. This is the solution for both of the equations of variables x and y.

STEP 3: PRACTICE/TESTING

1. **Write the code to find the arithmetic mean of the cut off marks of all male students.**

```
getwd()
data=read.csv("Studentdetails1.csv")
data
m = subset(data,GENDER=="Male")m
mean(m$CUTOFFMARKS)
```

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

A.M. of a set of observations is their sum divided by the number of observations

For ungrouped data:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{(x_1 + x_2 + \cdots + x_n)}{n}$$

For ungrouped frequency distribution with values $x_1, x_2, \ldots, x_n$ and corresponding frequencies $f_1, f_2, \ldots, f_n$:

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{n} f_i x_i = \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i}$$

For grouped frequency distribution,

$\bar{x} = \frac{1}{N}\sum_{i=1}^{n} f_i x_i = \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i}$ where x_i is the midpoint of the corresponding class.

RAW DATA:

```
A=c(……)
B=mean (A)
B
```

FREQUENCY DISTRIBUTION:

```
d=read. table (header=TRUE, text= "x        f
                                  ……..")
d1=rep (d$x, d$f)
A=mean (d1)
A
```

3. **Define median.**

Median is the value of the middle item when the items are arranged in ascending or descending order of magnitude. It is a positional average.

4. **Define mode.**

Mode is the value which occurs most frequently in a set of observations.

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

Standard deviation is the positive square root of the arithmetic mean of the squares of the deviations of deviations from their arithmetic mean.

For raw data, $\sigma = \sqrt{\frac{1}{n}\sum(x_i - \bar{x})^2} = \sqrt{\frac{1}{n}\sum x_i^2 - \left(\frac{\sum x_i}{n}\right)^2}$

For frequency distribution,

$\sigma = \sqrt{\frac{1}{N}\sum f_i(x_i - \bar{x})^2} = \sqrt{\frac{1}{N}\sum f_i x_i^2 - \left(\frac{\sum f_i x_i}{N}\right)^2}$ where $N = \sum f_i$

RAW DATA:

A=c(…….)
B=sd (A)
B

FREQUENCY DISTRIBUTION:

d=read. table (header=TRUE, text= “x f
…..”)
d1=rep (dx, df)
C=sd (d1)
C

KUMARAGURU COLLEGE OF TECHNOLOGY

LABORATORY MANUAL

Experiment Number: 4

Lab Code	**: U18MAI4201**
Lab	**: Probability and Statistics**
Course / Branch	**: B.E-CSE,ISE, B.Tech-IT**
Title of the Experiment	**: 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.

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

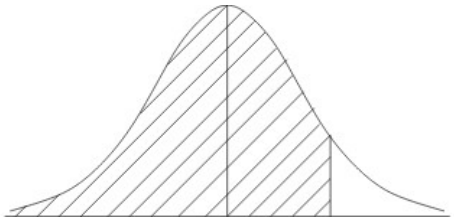
If a random variable X follows the normal distribution, then we write: $X \sim N(\mu, \sigma^2)$
The normal distribution with μ = 0 and σ = 1 is called the standard normal distribution, and is denoted as N(0,1).

Consider a normal distribution with mean μ and standard deviation σ

R-code for doing the Experiment:

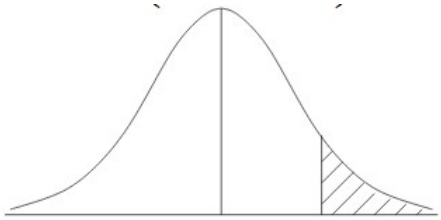
1.	To find $P(X < a) = P(-\infty < X < a)$ **R-code :** pnorm(a, mean = μ, sd = σ)
2.	To find $P(X > a) = P(a < X < \infty)$ R-code: pnorm(a, mean = μ, sd = σ, lower.tail = FALSE)
3.	To find $P(a < X < b)$ R-code: pnorm(b, mean = μ, sd = σ) - pnorm(a, mean = μ, sd = σ)

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



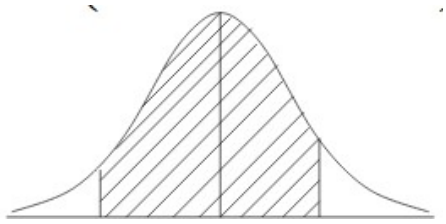
pnorm (a, mean = μ, sd = σ)

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)

Ans:

(i) pnorm(71, mean=68,sd=2.5)-pnorm(66,mean=68,sd=2.5)
[1] 0.6730749
Percentage = 67.30749%

(ii) pnorm(72.5,mean=68,sd=2.5)-pnorm(71.5,mean=68,sd=2.5)
[1] 0.04482634
Percentage = 4.482634%

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)

Ans:

(i) over=pnorm(700,mean=662,sd=32,lower.tail=FALSE)
print(over*1000)

[1] 117.5152

(ii) below=pnorm(650,mean=662,sd=32)

print(below*1000)

[1] 353.8302

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.**

Ans:

pnorm(2.52,mean=2.498,sd=0.012)-pnorm(2.48,mean=2.498,sd=0.012)
[1] 0.8998163
Percentage = 89.998163%

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.

Ans:

(i) pnorm(60,mean=42,sd=24,lower.tail=FALSE)
[1] 0.2266274

Number of children exceeding the score 60 = 0.2266274*1000 = 226.6 ≈ 227

(ii) pnorm(40,mean=42,sd=24)-pnorm(20,mean=42,sd=24)
[1] 0.2871346

Number of children with score lying between 20 and 40 = 0.2871346*1000
= 287.1346 ≈ 287

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*.

Ans:

(i)
```
one=pnorm(155,mean=151,sd=15)-pnorm(142,mean=151,sd=15)
print(one*500)
```

```
[1] 165.442
```

Number of students weigh between 142 and 155 lb ≈ 166

(ii)
```
two=pnorm(185,mean=151,sd=15,lower.tail=FALSE)
print(two*500)
```

```
[1] 5.852649
```

Number of students weigh more than 185 lb ≈ 6

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?

Ans:

(i)
```
pnorm(2000,mean=1500,sd=500,lower.tail=FALSE)
[1] 0.1586553
```

Percentage = 0.1586553*100 = **15.86 %**

(ii)
```
pnorm(1700,mean=1500,sd=500)-pnorm(1200,mean=1500,sd=500)
[1] 0.3811686
```

Percentage = 0.3811686*100 = **38.12 %**

STEP 3: PRACTICE/TESTING

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

A continuous random variable X follows normal distribution (or Gaussian distribution) then its p.d.f is

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \quad , -\infty < x < \infty, \quad -\infty < \mu < \infty, \quad \sigma > 0$$

The parameters are μ and σ where μ is the mean and σ is the standard deviation of the distribution.

2. **Define standard normal distribution.**

The Normal Distribution with mean = 0 and variance = 1. If X is a normal variate, then $z = \frac{X-\mu}{\sigma}$ is a standard normal variate. The p.d.f of the standard normal variate is

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

Area under the standard normal curve = 1

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

1. The graph of the distribution is bell shaped and is called the normal probability curve.
2. The curve is symmetrical about the ordinate at x = μ.
3. x – axis is an asymptote to the curve.
4. For the normal distribution, mean = median = mode.

KUMARAGURU COLLEGE OF TECHNOLOGY

LABORATORY MANUAL

Experiment Number: 2

Lab Code	**: U18MAI4201**
Lab	**: Probability and Statistics**
Course / Branch	**: B.E-CSE,ISE, B.Tech-IT**
Title of the Experiment	**: 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)
```

1. **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:

```
T1=c(1780, 1760, 1690, 1750, 1840, 1920, 1100, 1810, 1050, 1950)
t1m=mean(T1)
t1md=median(T1)
t1m
t1md
```

Output:

```
> t1m
[1]  1665

> t1md
[1] 1770
```

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:

```
T2=c(105.26,101.05,113.60,105.97,95.05,93.58,76.21,67.42,77.88,77.97,104.44,174.11)
t2m=mean(T2)
t2md=median(T2)
t2m
t2md
```

Output:

```
> t2m

[1] 99.37833


> t2md

[1] 98.05
```

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.

R-Code:

```
T3=c(1945.63, 1864.69, 1093.11, 1297.27, 918.15)
t3m=mean(T3)
t3md=median(T3)
t3m
t3md
```

Output:

```
> t3m
[1] 1423.77

> t3md
[1] 1297.27
```

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

R – Code:

```
d=read.table(header=TRUE,text="x            f
                                    1          5
                                    2          9
                                    3          12
                                    4          17
                                    5          14
                                    6          10
                                    7          6")
t4= rep(d$x, d$f)
multi.fun = function(fr)
{
  c(mean=mean(fr),sd=sd(fr))
}
multi.fun(t4)
```

Output:

```
     mean              sd
4.095890       1.668036
```

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.

R – Code:

```
d=read.table(header=TRUE,text="Miles         Villagers
                                2              38
                                4              104
                                6              140
                                8              78
                                10             48
                                13             42
                                14             28
                                16             24
                                18             16
                                20             2")
t5= rep(d$Miles, d$Villagers)
multi.fun = function(fr)
{
  c(mean=mean(fr),median=median(fr))
}
multi.fun(t5)
```

Output:

```
    mean      median
7.857692    6.000000
```

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

R – Code:

```
d=read.table(header=TRUE,text="Size        Frequency
                                6          3
                                7          6
                                8          9
                                9          13
                                10         8
                                11         5
                                12         4")
t6= rep(d$Size, d$Frequency)
multi.fun = function(fr)
{
  c(mean=mean(fr),sd=sd(fr))
}
multi.fun(t6)
```

Output:

```
      mean            sd
9.000000   1.624284
```

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

R – Code:

```
mode=function(x)
{
  ux= unique(x)
  ux[which.max(tabulate(match(x,ux)))]
}
T7=c(14.8,14.2,13.8,13.5,14.0,14.2,14.3,14.6,13.9,14.0,14.1,13.2,13.0,14.2,13.5,13.0,12.8,13.9,1
4.8,15.0,12.8,13.4,13.2,14.0,13.8,13.9,14.0,14.0,13.9,14.8)
c(mean=mean(T7),median=median(T7),mode=mode(T7))
```

Output:

```
    mean      median        mode
13.88667   13.95000   14.00000
```

KUMARAGURU COLLEGE OF TECHNOLOGY

LABORATORY MANUAL

Experiment Number: 5

Lab Code : **U18MAI4201**

Lab : **Probability and Statistics**

Course / Branch : **B.E-CSE,ISE, B.Tech-IT**

Title of the Experiment : **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 νd.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 $t = \frac{\overline{x}-\mu}{S/\sqrt{n}}$ follows t – distribution with n-1 degrees of freedom.

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

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

If $|t| \leq$ tabulated t, then H_0 is accepted and the difference between $\overline{x}$ 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 $t = \frac{\overline{x}-\overline{y}}{S\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim t_{n_1+n_2-2}$

where $\overline{x} = \frac{\sum x}{n_1}, \overline{y} = \frac{\sum y}{n_2}$,

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

or $S^2 = \frac{1}{n_1+n_2-2}[n_1 {s_1}^2 + n_2 {s_2}^2]$, where ${s_1}^2 = \frac{1}{n_1}\sum_i (x_i - \overline{x})^2, {s_2}^2 = \frac{1}{n_2}\sum_j \left(y_j - \overline{y}\right)^2$

(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 $\overline{x}$ 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 $t = \frac{\overline{d}}{S/\sqrt{n}}$ where $\overline{d} = \frac{1}{n}\sum_{i=1}^{n} d_i$ and $d_i = x_i - y_i, S^2 = \frac{1}{n-1}\sum\left(d_i - \overline{d}\right)^2$

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

$$Null Hypothesis: H_0: \mu = 66$$

$$\boldsymbol{Alternative Hypothesis: H_1: \mu \neq 66}$$

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: $\boldsymbol{H_0: \mu = 577}$

Alternate hypothesis: $H_1: \mu \neq 577$

R-code

```
x = c(578,572,570,568,572,570,570,572,596,584)

t.test(x,alternative="two.sided",mu=577)
```

Output:

One Sample t-test

data: x

t = -0.65408, df = 9, p-value = 0.5294

alternative hypothesis: true mean is not equal to 577

95 percent confidence interval:

568.9746 581.4254

sample estimates:

mean of x

575.2

Conclusion:
t -value = 0.65408 < 2.262. Hence we accept H_0 and we may conclude that the mean breaking stre ngth of the wire can be assumed as 577kg.

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: $\mu = 64$

Alternate hypothesis: $H_1: \mu > 64$

R-code:

```
x = c(70, 67, 62, 68, 61, 68, 70, 64, 64, 66)
t.test(x,alternative="greater",mu=64)
```

Output :

One Sample t-test

data: x
t = 2, df = 9, p-value = 0.03828
alternative hypothesis: true mean is greater than 64
95 percent confidence interval:
64.16689 Inf
sample estimates:
mean of x
66

Conclusion:
t -value = 2 > 1.833 . Hence we reject H_0 and we may conclude that the mean height is greater than 6 4 inches.

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 hypothesisH_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 hypothesisH_0: $\mu_1 = \mu_2$, ie, Students of both schools performed equally well.

Alternate hypothesisH_1: $\mu_1 > \mu_2$ie, Students of school A performed better than students of school B.

R-code:

```
A = c(76,68,70,43,94,68,33)

B = c(40,48,92,85,70,76,68,22)

t.test(A,B,alternative="greater",var.equal=TRUE)
```

Output:

```
        Two Sample t-test

data:  A and B
t = 0.16802, df = 13, p-value = 0.4346
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 -18.56956       Inf
sample estimates:
mean of x mean of y
 64.57143  62.62500
```

Conclusion:

t -value = 0.16802 < 1.771. Hence we accept H_0 ,we may conclude that there is no significant difference in the performance of the students of the two schools.

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: $\mu_1 = \mu_2$, ie, There is no significant difference between the means of the two samples.

Alternate hypothesis H_1: $\mu_1 \neq \mu_2$ ie, There is a significant difference between the means of the two samples.

R-code:

```
Samp1 = c(19,17,15,21,16,18,16,14)

Samp2 = c(15,14,15,19,15,18,16)

t.test(Samp1,Samp2,alternative="two.sided",var.equal=TRUE)
```

Output:

```
        Two Sample t-test


data:  Samp1 and Samp2

t = 0.93095, df = 13, p-value = 0.3688

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

 -1.320608  3.320608

sample estimates:

mean of x mean of y

      17       16
```

Conclusion:

t -value = 0.93095 < 2.160. Hence we accept H_0 ,we may conclude that there is no significant difference in the means of the two samples.

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: $\mu_1 = \mu_2$, ie, There is no significant difference between the two systems of feeding.

Alternate hypothesis H_1: $\mu_1 \neq \mu_2$ ie, There is a significant difference between the two systems of feeding.

R-code:

```
SysA=c(43,39,39,42,46,43,38,44,51,43)

SysB=c(37,35,34,41,39,37,37,40,48,36)

t.test(SysA,SysB,alternative="two.sided",paired=TRUE)
```

Output:

```
        Paired t-test


data:  SysA and SysB

t = 6.2644, df = 9, p-value = 0.0001471

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

 2.811113 5.988887

sample estimates:

mean of the differences

           4.4
```

Conclusion:

t -value = 6.2644 > 2.262. Hence we reject H_0 ,we may conclude that there is a significant difference between the two systems of feeding.

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: H_0: $\mu_1 = \mu_2$, ie, The employees have not been benefitted by the training.

Alternate hypothesis H_1: $\mu_1 < \mu_2$ ie, The employees have been benefitted by the training.

R-code:

```
Before=c(80,76,92,60,70,56,74,56,70,56)

After=c(84,70,96,80,70,52,84,72,72,50)

t.test(Before,After,alternative="less",paired=TRUE)
```

Output:

```
        Paired t-test


data:  Before and After

t = -1.4142, df = 9, p-value = 0.09547

alternative hypothesis: true difference in means is less than 0

95 percent confidence interval:

    -Inf 1.184826

sample estimates:

mean of the differences

                -4
```

Conclusion:

t -value = 1.4142 < 1.83. Hence we accept H_0 ,we may conclude that the employees have not been benefitted by the training.

STEP 3: PRACTICE/TESTING

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

   T=(x-y)/(s/sqrt(n))

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

   T=(x-y)/s(sqrt(1/n1+1/n2))

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

   T= d/(s/sqrt(n))

4. **Define level of significance.**

The significance level of an event is the probability that the event could have occurred by chance

KUMARAGURU COLLEGE OF TECHNOLOGY

LABORATORY MANUAL

Experiment Number: 6

Lab Code	**: U18MAI4201**
Lab	**: Probability and Statistics**
Course / Branch	**: B.E-CSE,ISE, B.Tech-IT**
Title of the Experiment/experiment	**: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 $F < F_{0.05}$, we accept 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: H_0: The population have the same variances.

Alternative Hypothesis: H_1: The population have different variances.

R Code:

```
Samp1=c(20,16,26,27,23,22,18,24,25,19)

Samp2=c(27,33,42,35,32,34,38,28,41,43,30,37)

var.test(Samp1,Samp2,ratio = 1,alternative = c("two.sided"),conf.level = 0.95)
```

Output:

```
        F test to compare two variances


data:  Samp1 and Samp2

F = 0.46709, num df = 9, denom df = 11, p-value = 0.2629

alternative hypothesis: true ratio of variances is not equal to 1

95 percent confidence interval:

 0.1301852 1.8272959

sample estimates:

ratio of variances

      0.4670913
```

Critical value of F for (9,11) d.f. is $F_{0.05} = 2.90$

Conclusion:

Since $F < F_{0.05}$, we accept the null hypothesis and we may conclude that the population hav e the same variances.

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: H_0: The population have the same variances.

Alternative Hypothesis: H_1: The population have different variances.

R Code:

```
Samp1=c(21,24,25,26,27)

Samp2=c(22,27,28,30,31,36)

var.test(Samp1,Samp2,ratio = 1,alternative = c("two.sided"),conf.level = 0.95)
```

Output:

```
        F test to compare two variances


data:  Samp1 and Samp2

F = 0.24537, num df = 4, denom df = 5, p-value = 0.1981

alternative hypothesis: true ratio of variances is not equal to 1

95 percent confidence interval:

 0.03321253 2.29776367

sample estimates:

ratio of variances

      0.2453704
```

Critical value of F for (4,5) d.f. is $F_{0.05} = 5.19$

Conclusion:

Since $F < F_{0.05}$, we accept the null hypothesis and we may conclude that the population hav e the same variances.

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: $\boldsymbol{H_0}$: $\sigma_1{}^2 = \sigma_2{}^2$

Alternative Hypothesis: $\boldsymbol{H_1}$: $\sigma_1{}^2 \neq \sigma_2{}^2$

R Code:

```
Sample1=c(9,11,13,11,15,9,12,14)

Sample2=c(10,12,10,14,9,8,10)

var.test(Sample1,Sample2,ratio = 1,alternative = c("two.sided"),conf.level = 0.95)
```

Output:

F test to compare two variances

data: Sample1 and Sample2

F = 1.2108, num df = 7, denom df = 6, p-value = 0.8315

alternative hypothesis: true ratio of variances is not equal to 1

95 percent confidence interval:

0.2125976 6.1978188

sample estimates:

ratio of variances

1.210843

Critical value of F for (7,6) d.f. is $F_{0.05} = 4.21$

Conclusion:
Since $F < F_{0.05}$, we accept the null hypothesis and we may conclude that the population hav e the same variances.

***t*-test:**

Null Hypothesis: H_0: $\mu_1 = \mu_2$,

Alternate hypothesis H_1: $\mu_1 \neq \mu_2$

R Code:

```
Sample1=c(9,11,13,11,15,9,12,14)

Sample2=c(10,12,10,14,9,8,10)

t.test(Sample1,Sample2,alternative="two.sided",var.equal=TRUE)
```

Output:

Two Sample t-test

data: Sample1 and Sample2

t = 1.2171, df = 13, p-value = 0.2452

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-1.024204 3.667061

sample estimates:

mean of x mean of y

11.75000 10.42857

Conclusion:

t -value = 1.2171 < 2.160 . Hence we accept H_0 ,we may conclude that the population means are same.

Final conclusion:

Since both the null hypothesis are accepted, we may conclude that the given samples have drawn from the same normal population.

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.

F – Test:

Null Hypothesis: H_0: ${\sigma_1}^2 = {\sigma_2}^2$ The two horses have the same running capacity in terms of variance of time taken.

Alternative Hypothesis: $\boldsymbol{H_1}$: ${\sigma_1}^2 \neq {\sigma_2}^2$ The two horses does not have the same running capacity in terms of variance of time taken.

R Code:

```
HorseA=c(28,30,32,33,33,29,34)

HorseB=c(29,30,30,24,27,29)

var.test(HorseA,HorseB,ratio = 1,alternative = c("two.sided"),conf.level = 0.95)
```

Output:

```
        F test to compare two variances


data:  HorseA and HorseB

F = 0.97604, num df = 6, denom df = 5, p-value = 0.9573

alternative hypothesis: true ratio of variances is not equal to 1

95 percent confidence interval:

 0.1398802 5.8441186

sample estimates:

ratio of variances

      0.9760426
```

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

Conclusion:

Since $F < F_{0.05}$, we accept the null hypothesis and we may conclude that the two horses have the same running capacity in terms of variance of time taken.

T - Test:

Null Hypothesis: $\boldsymbol{H_0}$**:** $\mu_1 = \mu_2$, The two horses have the same running capacity in terms of average of time taken.

Alternate hypothesis H_1: $\mu_1 \neq \mu_2$ The two horses does not have the same running capacity in terms of average of time taken.

R Code:

```
HorseA=c(28,30,32,33,33,29,34)

HorseB=c(29,30,30,24,27,29)

t.test(HorseA,HorseB,alternative="two.sided",var.equal=TRUE)
```

Output:

```
        Two Sample t-test


data:  HorseA and HorseB

t = 2.436, df = 11, p-value = 0.03306

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

 0.3009233 5.9371719

sample estimates:

mean of x mean of y

 31.28571  28.16667
```

Conclusion:

t -value = 2.436 > 2.201 . Hence we reject H_0 ,we may conclude that the two horses does not have the same running capacity in terms of average of time taken.

Final Conclusion:

In the t-test, the null hypothesis is rejected. So the two horses have the same running capacity only in terms of variance and not in terms of average of time taken.

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: $\boldsymbol{H_0}$: ${\sigma_1}^2 = {\sigma_2}^2$ the two samples have the same variance.

Alternative Hypothesis: $\boldsymbol{H_1}$: ${\sigma_1}^2 \neq {\sigma_2}^2$ the two samples have different variance.

R Code:

```
Samp1=c(60,65,71,74,76,82,85,87)
Samp2=c(61,66,67,85,78,63,85,86,88,91)
var.test(Samp1,Samp2,ratio = 1,alternative = c("two.sided"),conf.level = 0.95)
```

Output:

```
        F test to compare two variances

data:  Samp1 and Samp2
F = 0.68143, num df = 7, denom df = 9, p-value = 0.6271
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.1623591 3.2866779
sample estimates:
ratio of variances
         0.6814286
```

Critical value of F for (7,9) d.f. is $F_{0.05} = 3.29$

Conclusion:

Since $F < F_{0.05}$, we accept the null hypothesis and we may conclude that the two samples have the same variance at 5 % level.

STEP 3: PRACTICE/TESTING

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

The main use of F distribution is to chech whether two independent samples have been drawn for the same variance or if two independent estimates of the population variance are homogeneous or not, since it is often desirable to compare two variance rather than two averages.

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

1)F- distribution is positively skewed.

2)Value of F lies between 0 and ∞.

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

t-test is used to test if two sample have the same mean. The assumptions are that they are samples from normal distribution. F-test is used to test if two sample have the same variance.

Register Number:……..

KUMARAGURU
college of technology
character is life

WORKSHEET -2

RAW DATA:

➢ MEAN:

```
A=c(1,2,3,4,4,5,6,6,6,1)
B=mean (A)
B
```

➢ MEDIAN:

```
A=c(1,2,3,4,4,5,6,6,6,1)
B=median (A)
B
```

➢ MODE:

```
A=c(1,2,3,4,4,5,6,6,6,1)
t=table (as.vector (A))
t
names(t)[t==max(t)]
```

➢ STANDARD DEVIATION:

```
A=c(1,2,3,4,4,5,6,6,6,1)
B=sd (A)
B
```

FREQUENCY DISTRIBUTION:

```
d=read. table (header=TRUE, text= "x        f
                                  5          15
                                 15          20
                                 25          30
                                 35          20")
d1=rep (d$x, d$f)
A=mean (d1)
B=median (d1)
C=sd (d1)
A
B
C
```

WORKSHEET-5

t-test:

➢ To test hypothesis about population mean:

(a) For a two-tailed test

```
x = c(63,63,66,67,68,69,70,70,71,71,71)

t.test(x,alternative="two.sided",mu=66)
```

(b) **For a one-tailed test**

x = c($a_1, a_2, ...a_N$)

t.test(x,alternative="less"/"greater",mu= μ)

➢ 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)

eg:

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)

- **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)

eg:

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)

NOTE:

If ‘p’ value is greater than 0.05, we accept H_0 and if it is less than 0.05, we reject H_0.