```
## [1] -1 -1 -1 -1

#constant multiplication
5*a

#product
a*b

#division

a/b

# character object is used to represent string values in R
X=as.character(5.2)
X

#Concatenation of strings
paste("Baa", "Baa", "Black", "Sheep")
```

\

The purpose of this experiment is to learn the different alignment of data set and various graphical representations in R

```
#creating a vector empid
empid=c(1,2,3,4,5,6,7,8,9,10,11,12,13,14,15)
empid
# creating a vector age
age=c(30,37,45,32,50,60,35,32,34,43,32,30,43,50,60)
Age
# creating a vector gender
gender=c(0,1,0,1,1,1,0,0,1,0,0,1,1,0,0)
gender

# creating a vector status
status=c(1,1,2,2,1,1,1,2,2,1,2,1,2,1,2)
Status
# reating a data frame (Combining vectors)
empinfo=data.frame(empid,age,gender,status)
empinfo
```

```
# Extract
```

Aim: To understand the simple correlation and linear regression with computation and interpretation

```
 male data
male=subset(empinfo,empinfo$gender=="male")
male


# Extract female data
female=subset(empinfo, empinfo$gender=='female')
female

# summary statistics for empinfo data
summary(empinfo)
```

```
# labeling character to numeric
empinfo$gender=factor(empinfo$gender,labels=c("male","female"))
empinfo$gender


empinfo$status=factor(empinfo$status,labels=c("staff","faculty"))
empinfo$statusempinfo
```

```
# summary statistics of male, female and age
summary(male)

summary(female)
```

```
summary(age)

# creating table (one-way)
table1=table(empinfo$gender)
table1

table2=table(empinfo$status)
table2


# creating table (two-way)
table3=table(empinfo$gender, empinfo$status)
table3
```

```r
# Graphical representation (Pie chart)
pie(table1)
```

```r
# Graphical representation (Box plot)
boxplot(empinfo$age~empinfo$status,col=c('red','blue'))
# Graphical representation (Bar plot)
barplot(table3,beside=T,xlim=c(1,15),ylim=c(0,5),col=c("blue", "red"))
legend("topright",legend=rownames(table3),fill=c('blue','red'),bty="n")


# Problem-1
# Import the inbuilt data set "cars"
data=cars
data
```

```
##    speed dist
## 1      4    2
## 2      4   10
```

```r
# Variance of "speed"
v1=var(data$speed)
v1
```

\

```r
# Variance of "dist"
v2=var(data$dist)
v2

# Covariance between "speed" and "dist"
covariance=cov(data$speed,data$dist)
covariance


#or
covariance=var(data$speed,data$dist)
covariance

# correlation coefficient using Pearson's formula
corr=covariance/(sd(data$speed)*sd(data$dist))
corr


# or
corr=cor(data$speed,data$dist)
corr

# Test for association between paired samples
cor.test(data$speed,data$dist)

cor.test(data$speed,data$dist,method="pearson")


cor.test(data$speed,data$dist,method="spearman")



# Visualize the samples
plot(data$speed,data$dist)
# Linear Regression model of "speed" with respect to "dist"
regression1=lm(data$speed~data$dist)
```

```
regression1
```

```
# Visualize linear regression line
abline(regression1)
summary(regression1)


# Linear Regression model of "dist" with respect to "speed"
regression2=lm(data$dist~data$speed)
regression2

abline(regression2)
```

Problem:-

The body weight and the BMI of 12 school going children are given in the following table

Weight	***15***	***26***	***27***	***25***	***25.5***	***27***	***32***	***18***	***22***	***20***	***26***	***24***
BMI	***13.35***	***16.12***	***16.74***	***16.00***	***13.59***	***15.73***	***15.65***	***13.85***	***16.07***	***12.8***	***13.65***	***14.42***

Let us fit a simple regression model BMI on weight and examine the results.

```r
#Problem-2
weight=c(15,26,27,2,25.5,27,32,18,22,20,26,24)
weight

bmi=c(133.35,16.1,16.74,16,13.59,15.73,15.65,13.85,16.07,12.8,13.65,14.42)
bmi


cor(weight,bmi)


mode1<-lm(bmi~weight)
summary.lm(mode1)
```

Problem 1: *The sale of a Product in lakhs of rupees(Y) is expected to be influenced by two variables namely the advertising expenditure X1 (in'OOORs) and the number of sales persons(X2) in a region. Sample data on 8 Regions of a state has given the following results*

Area	Y	X1	X2
1	110	30	11
2	80	40	10
3	70	20	7
4	120	50	15
5	150	60	19
6	90	40	12
7	70	20	8
8	120	60	14

```r
# Input the variables
Y=c(110,80,70,120,150,90,70,120)
Y

X1=c(30,40,20,50,60,40,20,60)
X1

X2=c(11,10,7,15,19,12,8,14)
X2


# Linear regression model of Y on X1 and X2
RegModel=lm(Y~X1+X2)
RegModel









# Summary of the data
summary(RegModel)2)
```

```
## [16] 10.4 14.7 32.4 30.4 33.9 21.5 15.5 15.2 13.3 19.2 27.3 26.0 30.4 15.8
# install.packages("scatterplot3d")
library(scatterplot3d)
# Plot the data set
scatterplot3d(Y,X1,X2)
Y


275.8

350.0




#Problem 2
data=mtcars
data

X=mtcars$mpg
X




summary(RegModel)




                                    Max




library(scatterplot3d)
graph=scatterplot3d(X,Y,Z)
# Visualize the plane
graph$plane3d(RegModel)
```

```r
# number of trials
n=4
n

#probability of success
p=0.02
p


# (i) probability of getting exactly 2 heads
dbinom(2,n,p)
```

```r
# (ii) probability of getting atleast 2 heads
sum(dbinom(2:4,n,p))


#or
1-pbinom(1,n,p)


# (iii) probability of getting atmost 2 heads
sum(dbinom(0:2,n,p))

# or
pbinom(2,n,p)

#(iv) Expectation of x
x=0:n
px=dbinom(x,n,p)
Ex=weighted.mean(x,px)
Ex

# (v) variance of x
Varx=weighted.mean(x*x,px)-(weighted.mean(x ,px))^2
Varx

# (vi) Visualize probability distribution
plot(x,px,type="h",xlab="values of x",ylab="Probability distribution of
x",main="Binomial distribution")
```

Problem:
A manufacturer of pins knows that 2% of his products are defective. If he sells pins in boxes of 20 and find the number of boxes containing (i) at least 2 defective (ii) exactly 2 defective (iii) at most 2 defective pins in a consignment of 1000 boxes (iv) plot the distribution (v) E(x) (vi) Variance of X?

Codes and Results:

```
#Poisson Distribution
# number of trails
m=20
m


# probability of success
ps=0.02
# poisson parameter
lambda=m*ps
lambda
```

```
#at least 2 defectives
p1=sum(dpois(2:m,lambda))
p1


# (i) number of boxes containing at least 2 defectives
round(1000*p1)

#exactly 2 defectives
p2=dpois(2,lambda)
p2

# (ii) number of boxes containing exactly 2 defectives
round(1000*p2)

#at most 2 defectives
p3=sum(dpois(0:2, lambda))
p3
# (iii) number of boxes containing at most 2 defectives
round(1000*p3)

# (iv) plot the distribution
x1=0:m
px1=dpois(x1,lambda)
plot(x1,px1,type="h",xlab="values of x",ylab="Probability distribution of
x",main="Poisson distribution")
```

```
#(v) E(x)
Ex1=weighted.mean(x1,px1)
Ex1


# (vi) variance of x
Varx1=weighted.mean(x1*x1,px1)-(weighted.mean(x1 ,px1))^2
Varx1
```

Problem:

A company finds that the time taken by one of its engineers to complete or repair job has a normal distribution with mean 20 minutes and S.D 5 minutes. State what proportion of jobs take:

i. Less than 15 minutes
ii. More than 25 minutes
iii. Between 15 and 25 minutes
iv. Plot the distribution
v. Table the distribution

```
# Generating the data x
x=seq(0,40)
x
# find the density function of x
y=dnorm(x,mean=20,sd=5)
y




# plot the normal distribution curve
plot(x,y,type='l')
# Proportion of jobs take less than 15 minutes
p1=pnorm(15,mean=20,sd=5)
p1
```

```
x2=seq(0,15)
x2

y2=dnorm(x2,mean=20,sd=5)
y2


polygon(c(0,x2,15),c(0,y2,0),col='yellow')

#Proportion of jobs take more than 25 minutes
p2=pnorm(40,mean=20,sd=5)-pnorm(25,mean=20,sd=5)
p2


x1=seq(25,40)
x1

y1=dnorm(x1,mean=20,sd=5)
y1


polygon(c(25,x1,40),c(0,y1,0),col='red')

#Proportion of jobs take between 15 and 25 minutes
p3=pnorm(25,mean=20,sd=5)-pnorm(15,mean=20,sd=5)
p3

x3=seq(15,25)
x3

y3=dnorm(x3,mean=20,sd=5)
y3

polygon(c(15,x3,25),c(0,y3,0),col='green')
```

```
# Probability distribution
data.frame(p1,p2,p3)

            \
```

Problem

Suppose the mean weight of King Penguins found in an Antarctic colony last year was 15.4 kg. In a sample of 35 penguins same time this year in the same colony, the mean penguin weight is 14.6 kg. Assume the population standard deviation is 2.5 kg. At 0 .05 significance level, can we reject the null hypothesis that the mean penguin weight does not differ from last year?

Codes and Results:

```
# Input the sample mean
xbar=14.6
xbar

# Input the population mean
mu0=15.4
mu0
```

```r
# Input the standard deviation
sigma=2.5
sigma


# Input the sample size
n=35
n


# Test Statistic
z=(xbar-mu0)/(sigma/sqrt(n))
z
```

```
## [1] -1.893146
```

```r
# Level of significance
alpha=0.05
alpha
```

```
## [1] 0.05
```

```r
# Two-tailed critical value
zhalfalpha=qnorm(1-(alpha/2))
zhalfalpha
```

```
## [1] 1.959964
```

```r
c(-zhalfalpha,zhalfalpha)
```

```
## [1] -1.959964  1.959964
```

```r
# To find p-value
pval=2*pnorm(z)
pval
```

```
## [1] 0.05833852
```

```r
# conclusion
if(pval>alpha){print("Accept Null hypothesis")} else{print("Reject Null
hypothesis")}
```

```
## [1] "Accept Null hypothesis"
```

Problem:
The fatality rate of typhoid patients is believed to be 17.26%. In a certain year 640 patients suffering from typhoid were treated in a metropolitan hospital and only 63 patients died. Can you consider the hospital efficient?

Code and Results:

```
# Input the data
# Size of the sample
n=640
n
```

```
## [1] 640
```

```
# Sample proportion
Sprop=63/n
Sprop
```

```
## [1] 0.0984375
```

```
# Population proportion
Pprop=0.1726
Pprop
```

```
## [1] 0.1726
```

```
# probability of failure
q=1-Pprop
q
```

```
## [1] 0.8274
```

```
# test statistic
z=(Sprop-Pprop)/sqrt(Pprop*q/n)
z
```

```
## [1] -4.964736
```

```
#critical value
E=qnorm(.975)
```

```r
# critical region
c(-E,E)
```

```
## [1] -1.959964  1.959964
```

```r
# confidence interval
Sprop+c(-E,E)*sqrt(Pprop*(1-Pprop)/n)
```

```
## [1] 0.06915985 0.12771515
```

```r
# Conclusion
if(z>-E && z<E){print("Hospital is not efficient")} else{print("Hospital is
efficient")}
```

```
## [1] "Hospital is efficient"
```

Codes and Results:

```
# Input the sample mean
xbar=20
xbar

## [1] 20

ybar=15
ybar

## [1] 15

# Input the standard deviation
sigma=4
sigma

## [1] 4
```

```r
# Input the sample size
n1=500
n1
```

```
## [1] 500
```

```r
n2=400
n2
```

```
## [1] 400
```

```r
# Test Statistic
z=(xbar-ybar)/(sigma*sqrt((1/n1)+(1/n2)))
z
```

```
## [1] 18.6339
```

```r
# Level of significance
alpha=0.05
alpha
```

```
## [1] 0.05
```

```r
# Two-tailed critical value
zalpha=qnorm(1-(alpha/2))
zalpha
```

```
## [1] 1.959964
```

```r
# conclusion
if(z<=zalpha){print("Accept Null hypothesis")} else{print("Reject Null
hypothesis")}
```

```
## [1] "Reject Null hypothesis"
```

Problem:

In a large city A, 20% of a random sample of 900 schools boys had a slight physical defect. In another large city B, 18.5% of a random sample of 1600 school boys had the same defect. Is the difference between the proportions significant?

Code and Results:

```r
# Input the sample proportions
p1=0.20
p1
```

```
## [1] 0.2
```

```r
p2=0.185
p2
```

```
## [1] 0.185
```

```r
# Input the sample size
n1=900
n1
```

```
## [1] 900
```

```r
n2=1600
n2
```

```
## [1] 1600
```

```r
# To find approximate population proportion
P=(n1*p1+n2*p2)/(n1+n2)
P
```

```
## [1] 0.1904
```

```r
Q=1-P
# Test Statistic
z=(p1-p2)/sqrt(P*Q*sqrt((1/n1)+(1/n2)))
z
```

```
## [1] 0.1871665
```

```r
# Level of significance
alpha=0.05
alpha
```

```
## [1] 0.05
```

```r
# Two-tailed critical value
zalpha=qnorm(1-(alpha/2))
zalpha
```

```
## [1] 1.959964
```

```r
# conclusion
if(z<=zalpha){print("Accept Null hypothesis")} else{print("Reject Null
hypothesis")}
```

```
## [1] "Accept Null hypothesis"
```

Problem: 1 (Student's t-test)

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	20

Is the difference between the sample means significant?

Code and Results:

```
# Problem 1

# input the data
sample1=c(19,17,15,21,16,18,16,14)
sample1

## [1] 19 17 15 21 16 18 16 14

sample2=c(15,14,15,19,15,18,16,20)
sample2

## [1] 15 14 15 19 15 18 16 20

# output using t-distribution
t=t.test(sample1,sample2)
t


# test-statistic
cv=t$statistic
cv

#critical value
tv=qt(0.975,14)
tv

## [1] 2.144787

#conclusion
if(cv <= tv){print("Accept Ho")} else{print("Reject Ho")}

## [1] "Accept Ho"
```

Problem: 2 (Paired t-test)

The following data relate to the marks obtained by 10 students in two test, one held at the beginning of a year and the other at the end of the year after intensive coaching. Do the data indicate that the students have got benefited by coaching?

Test 1	19	17	15	21	16	18	16	14	19	20
Test 2	15	14	15	19	15	18	16	20	22	19

```
# Problem 2

#Paired- t-test
# input the data
test1=c(19,17,15,21,16,18,16,14,19,20)
test1

## [1] 19 17 15 21 16 18 16 14 19 20

test2=c(15,14,15,19,15,18,16,20,22,19)
test2

##  [1] 15 14 15 19 15 18 16 20 22 19

t=t.test(sample1,sample2,paired=TRUE)
t

# level of significance
alpha=0.05
# p-value
tv=t$p.value
tv

## [1] 0.6542055

# conclusion
if(tv >alpha){print("Accept Ho")} else{print("Reject Ho")}

## [1] "Accept Ho
```

Problem: 3 (F-test)

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	20

Is the difference between the sample means significant?

Codes and Results:

```
# Problem 3

# Variance test or F-test
sample1=c(19,17,15,21,16,18,16,14)
sample1

## [1] 19 17 15 21 16 18 16 14

sample2=c(15,14,15,19,15,18,16,20)
sample2

## [1] 15 14 15 19 15 18 16 20

# output using t-distribution
f=var.test(sample1,sample2)
f
```

```r
# test-statistic
cv=f$statistic
cv
```

```
##        F
## 1.058824
```

```r
#critical value
tv=qf(0.95,7,7)
tv
```

```
## [1] 3.787044
```

```r
#conclusion
if(cv <= tv){print("Accept Ho")} else{print("Reject Ho")}
```

\

Problem: 1

Five coins are tossed 256 times. The number of heads observed by binomial distribution is given below. Examine if the coins are unbiased by employing chi-square goodness of fit.

No. of heads	0	1	2	3	4	5
Frequency	5	35	75	84	45	12

Codes and Results:

```
# Problem : 1

# Goodness of fit

# Number of coins
n=5
n


# level of significance
alpha=0.05
alpha

## [1] 0.05

N=256 # Total number of tosses
N

## [1] 256

P = 0.5  # probability of getting head
P

## [1] 0.5

x = c(0:n);x

## [1] 0 1 2 3 4 5

obf = c(5,35,75,84,45,12)# observed frequencies
obf

## [1]  5 35 75 84 45 12

exf = (dbinom(x,n,P)*256) # expected frequencies
exf

## [1]  8 40 80 80 40  8

# check the condition if the observed and expected frequencies sum are equal
sum(obf)

## [1] 256

sum(exf)

## [1] 256

# output using Chisq-distribution
chisq<-sum((obf-exf)^2/exf)
cv = chisq;cv
```

```
## [1] 4.8875
```

```
# critical value using Chisq-distribution
tv = qchisq(1-alpha,n);tv
```

```
## [1] 11.0705
```

```
# Hypothesis conclusion
if(cv <= tv){print("Accept H0/Fit is good")} else{print("Reject H0/Fit is not good")}
```

```
## [1] "Accept H0/Fit is good"
```

Problem: 2

From the following information state whether the condition of the child is associated with the condition of the house.

Condition of child	**Condition of house Clean**	**Condition of house dirty**
Clean	69	51
Fairly clean	81	20
dirty	35	44

Codes and Results:

```
# Problem : 2
```

```
# Independent of attributes
# Input the data
data<-matrix(c(69,51,81,20,35,44),ncol=2,byrow=T)
data
```

```
# number of data
l=length(data);l
```

```
## [1] 6
```

```
# output by Chisq-distribution
cv=chisq.test(data)
cv
```

```r
# Hypothesis conclusion
if(cv >alpha){print("Attributes are independent")} else{print("Attributes are
not independent")}
```

```
## [1] "Attributes are not independent"
```

.

Problem:
A car rental agency, which uses 5 different brands of tyres in the process of deciding the brand of tyre to purchase as standard equipment for its fleet, finds that each of 5 tyres of each brand last the following number of kilometres (in thousands):

A	B	C	D	E
36	46	35	45	41
37	39	42	36	39
42	35	37	39	37
38	37	43	35	35
47	43	38	32	38

Test the hypothesis that the five brands have almost the same average life.

Code and Results:

```
#One-way  ANOVA
# Types of tyres
A=c(36,37,42,38,47)
B=c(46,39,35,37,43)
C=c(35,42,37,43,38)
D=c(45,36,39,35,32)
E=c(41,39,37,35,38)
group<-data.frame(cbind(A,B,C,D,E))
group
```

```
summary(group)
```

```
# stack vector from data frame
stgr<-stack(group);stgr


# completely randomized design
crd<-aov(values~ind,data=stgr)
# ANOVA table
summary(crd)

# Visualization of data
boxplot(group, ylab="Average life of tyres in kilometers",main="Brands of Tyres")
```

Problem:

Code and Results:

```
#Monthly sales of States
StateA=c(6,5,3,8)
StateA

## [1] 6 5 3 8

StateB=c(8,9,6,5)
StateB

## [1] 8 9 6 5

StateC=c(10,7,8,7)
StateC

## [1] 10  7  8  7

#frame the data set
Group<-data.frame(cbind(StateA,StateB,StateC))
Group

g=c("Salesman1","Salesman2","Salesman3","Salesman4")
g

## [1] "Salesman1" "Salesman2" "Salesman3" "Salesman4"

# number of columns
k=ncol(Group)
k

## [1] 3

# number of rows
n=nrow(Group)
n

## [1] 4

# Generate factor levels of States
States=gl(k,1,n*k,factor(f))
States


# Generate factor levels of Salesmen
Salesmen=gl(n,k,n*k,factor(g))
Salesmen


# ANOVA table
anova=aov(Sales ~ States + Salesmen)
summary(anova)
```

Preoblem

Perform Latin Square Design for the following.

Consider analyzing the productivity of five kinds of manure, five kinds of cultivation, and five

kinds of crops. As follows, the data are organized in a Latin Square format:

	cultP	cultQ	cultR	cultS	cultT
manure1	"P42"	"R47"	"Q55"	"S51"	"T44"
manure2	"T45"	"Q54"	"R52"	"P44"	"S50"
manure3	"R41"	"P46"	"DS7"	"T47"	"Q48"
manure4	"Q56"	"S52"	"T49"	"R50"	"P43"
manure5	"S47"	"T49"	"P45"	"Q54"	"R46"

The three factors are: manure (manure1:5), cultivation (cultP:T), crop(P:T).

Codes and Results:

```
#creating dataframes in R
manure=c(rep("manure1",1), rep("manure2",1), rep("manure3",1),
rep("manure4",1), rep("manure5",1))
cultivation=c(rep("cultP",5), rep("cultQ",5), rep("cultR",5), rep("cultS",5),
rep("cultT",5))
```

```
crop=c("P","T","R","Q","S", "R","Q","P","S","T", "Q","R","S","T","P",
"S","P","T","R","Q", "T","S","Q","P","R")
freq=c(42,45,41,56,47, 47,54,46,52,49, 55,52,57,49,45, 51,44,47,50,54,
44,50,48,43,46)
data=data.frame(cultivation,manure,crop,freq)
data
```

```
## Analysis of Variance Table
##
## Response: freq
##             Df Sum Sq Mean Sq F value   Pr(>F)
## manure       4  17.76   4.440  0.7967 0.549839
## cultivation  4 109.36  27.340  4.9055 0.014105 *
## crop         4 286.16  71.540 12.8361 0.000271 ***
## Residuals   12  66.88   5.573
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Conclusion: The problems on ANOVA have been executed using R

```
#recreating the original table, using the matrix function
matrix(data$crop,5,5)


matrix(data$freq,5,5)

#creating the anova table
fit=lm(freq~manure+cultivation+crop,data)
anova(fit)
```