MAT2001 Statistics for Engineers

LAB Manual

Please share your suggestions and feedback to

Dr. B JAGANATHAN
R-LAB COORDINATOR
MATHEMATICS DIVISION (SAS)
VIT- CHENNAI
jaganathan.b@vit.ac.in

Experiment No.	Title
1	**Introduction to R and Basic Commands**
2	**Computation of Tables and Graphs-Summary Statistics**
3	**Random Variable and Probability Distributions**
4	**Discrete and Contributions Probability Distributions**
5	**Correlation and Regression**
6	**Multiple Linear Regression**
7	**Testing of Hypothesis- I (Z test)**
8	**Testing of Hypothesis – II(t, F, Chi-square)**
9	**Completely Randomized Design**
10	**Randomized Block Design**

About R Language

- ✓ R is a computer language for carrying out statistical computations.
- ✓ R is Free Software, and runs on a variety of platforms
- ✓ Command-line execution based on function calls.
- ✓ Workspace containing data and functions.
- ✓ Extensible with user functions.
- ✓ Graphics devices.
- ✓ R packages can contain not only code, but also other resources like documentation and sample data sets.

- ✓ Well-defined format that ensures easy installation, a basic standard of documentation, and enhances portability and reliability.
- ✓ The basic mode of interaction is 'read – evaluate – print'.
- ✓ The R Project is an international collaboration of researchers in statistical computing.
- ✓ There are roughly 20 members of the "R Core Team" who maintain and enhance R.
- ✓ Releases of the R environment are made through the CRAN (comprehensive R archive network) twice per year.
- ✓ The software is released under a "free software" license, which makes it possible for anyone to download and use it.
- ✓ There are over 3500 extension packages that have been contributed to CRAN.
- ✓ R is a computer language which is processed by a special program called an interpreter. This program reads and evaluates R language expressions, and prints the values determined for the expressions.

Installation

- ✓ R can be downloaded from one of the mirror sites in http://cran.r-project.org/mirrors.html. You should pick your nearest location.
- ✓ Download R-studio 1.1.383 for Windows/Linux from Google search.

Using External Data

- ✓ R offers plenty of options for loading external data, including Excel, Minitab, SAS and SPSS files.

1. Basic Concepts in R, Understanding Data types, Importing/Exporting data

- ✓ After R is started, there is a console awaiting for input. At the prompt (>), you can enter numbers and perform calculations.

```
> 2+3
[1] 5
> 100+200+300
[1] 600
```

- ✓ **Functions :** R functions are invoked by its name, followed by the parenthesis and arguments. The function c is used to combine three numeric values into a vector

```
> c(1,2,3)
```

[1] 1 2 3

> c(100,200,300)

[1] 100 200 300

- ✓ All text after the pound sign "#" within the same line is considered a comment.

```
> 5                    # type 5 at the prompt
[1] 5                  # here 5 is returned
> 3 + 4          # adding two numbers
[1] 7
> 5^3                  # will compute 5^3
[1] 125
> pi                   # pi value
[1] 3.141593
> 1 + 2 * 3            # Normal arithmetic rules apply
[1] 7
```

- ✓ R's basic operators have the following precedence (listed in highest-to-lowest order)

^	exponentiation
- +	unary minus and plus
:	sequence operator
%/% %%	integer division, remainder
* /	multiplication, division
+ -	addition, subtraction

Example:-

>2^3^2	**> (2^3)^2**	**> 2^(3^2)**	**> sqrt(2)**
[1] 512	**[1] 64**	**[1] 512**	**[1] 1.414214**
> log(10)	**> log10(10)**	**> sin(1)**	**> 4*atan(1)**
[1] 2.302585	**[1] 1**	**[1] 0.841471**	**[1] 3.141593**

- ✓ The expression n1:n2, generates the sequence of integers from n1 to n2.

```
> 1:15    #print the numbers 1 to 15

[1]  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15

> 5:-5     # print the numbers 5 to -5

 [1]  5  4  3  2  1  0 -1 -2 -3 -4 -5
```

- ✓ NA is used to indicate that a value is missing or not available. Any arithmetic expression which contains NA will produce NA as a result.

> 1/0	**> Inf-Inf**	**> sqrt(-1)**	**> 1+sin(NA)**
[1] Inf	**[1] NaN**	**[1] NaN**	**[1] NA**

Assignment:-

- ✓ Values are stored by assigning them a name**.**
- ➢ The following statements all store the value 18 under the name.(= or -> or <-)

>x = 18 >x <- 18 >18 -> x

- ➢ Variables can be used in expressions in the same way as numbers.

```
> x=22
> x=x+25
> x
[1] 47
```

- ➢ Individual values can be combined into a vector by using the c function.

```
> x=c(1,2,3,4)
> x
[1] 1 2 3 4
```

Character:-

- ✓ A character object is used to represent string values in R**.**
- ➢ as.character () function is used to convert objects into character values:

```
> x=as.character(4.58)
> x
[1] "4.58"
```

- ➢ Strings can be concatenated by using paste function.

```
> paste("First", "Second", "Third")
[1] "First Second Third"
>paste("First", "Second", "Third", sep = ":")
[1] "First:Second:Third"

> fname = "Sri";  lname ="Ram"
> paste(fname)
> paste(fname,lname)
[1] "Sri Ram"
```

Vector Arithmetic:-

Arithmetic operations of vectors are performed member wise.

```
> a = c(1, 3, 5, 7)
> b = c(1, 2, 4, 8)
```

If we add a and b, the sum would be a vector whose members are the sum of the corresponding members from a and b.

```
> a+b
[1]  2  5  9 15
```

If we multiply a by 5, we get a vector with each of its members multiplied by 5.

```
> 5*a
[1]  5 15 25 35
```

Similarly for subtraction, multiplication and division, we get new vectors via member wise operations.

```
> a-b
[1]  0  1  1 -1
> a*b
[1]  1  6 20 56
> a/b
[1] 1.000 1.500 1.250 0.875
>  a=c(1,2,3,4)
> 2*a+1
[1] 3 5 7 9
```

If two vectors are of unequal length, the shorter one will be recycled in order to match the longer vector

```
> u=c(10,20,30)
> v=c(1,2,3,4,5,6,7,8,9)
> u+v
[1] 11 22 33 14 25 36 17 28 39
```

Data Frame:-

A **data frame** is used for storing data tables. It is a list of vectors of equal length. For example, the following variable df is a data frame containing three vectors n, s, b.

```
n = c(2, 3, 5)
> s = c("aa", "bb", "cc")
> b = c(TRUE, FALSE, TRUE)
> df = data.frame(n, s, b)          # df is a data frame
> df
  n  s     b
1 2 aa  TRUE
2 3 bb FALSE
3 5 cc  TRUE
```

For example, here is a built-in data frame in R, called mtcars.

```
> mtcars[1, 2]                      # first row, second column
[1] 6

> mtcars["Mazda RX4", "cyl"]        # using the row and column names
```

[1] 6

> nrow(mtcars) # number of data rows

[1] 32

> ncol(mtcars) # number of columns

[1] 11

Preview: Instead of printing out the entire data, it is often desirable to preview it with the head function beforehand.

> head(mtcars)

Data Import:-

CSV File: The sample data can also be in comma separated values (CSV) format. The first row of the data file should contain the column names instead of the actual data.

Important Note:

Enter the following data(or any data) in Excel sheet and save it as CSV file.

col1	col2	col3
34	23	76
56	54	43
76	34	24
54	76	67
32	24	54

Code:- For example

>mydata=read.csv ("C:\\Users\\admin\\Desktop\\mokesh\\workdata.csv")
select your file based on your path (or)
>mydata=read.csv (file.choose())

2. Computation of Tables and Graphs -Summary Statistics

Aim: To represent the various types of data using tabulation and graphical representation

Computation of tables and graphs-summary statistics for employee data:

Creating vector:-

>empid=c(1,2,3,4,5,6,7,8,9,10,11,12,13,14,15) #creating a vector empid

> empid

[1] 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15

> age=c(30,37,45,32,50,60,35,32,34,43,32,30,43,50,60) # creating a vector age

> age

[1] 30 37 45 32 50 60 35 32 34 43 32 30 43 50 60

> sex=c(0,1,0,1,1,1,0,0,1,0,0,1,1,0,0)

> sex

[1] 0 1 0 1 1 1 0 0 1 0 0 1 1 0 0

> status=c(1,1,2,2,1,1,1,2,2,1,2,1,2,1,2)

> status

[1] 1 1 2 2 1 1 1 2 2 1 2 1 2 1 2

Creating a data frame (Combining vectors):

> empinfo=data.frame(empid,age,sex,status)

> empinfo

	empid	age	sex	status
1	1	30	0	1
2	2	37	1	1
3	3	45	0	2
4	4	32	1	2
5	5	50	1	1
6	6	60	1	1
7	7	35	0	1
8	8	32	0	2
9	9	34	1	2
10	10	43	0	1
11	11	32	0	2
12	12	30	1	1
13	13	43	1	2
14	14	50	0	1
15	15	60	0	2

> empinfo$sex=factor(empinfo$sex,labels=c("male","female"))

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

>empinfo

```
   empid age    sex  status
1      1  30   male   staff
2      2  37 female   staff
3      3  45   male faculty
4      4  32 female faculty
5      5  50 female   staff
6      6  60 female   staff
7      7  35   male   staff
8      8  32   male faculty
9      9  34 female faculty
10    10  43   male   staff
11    11  32   male faculty
12    12  30 female   staff
13    13  43 female faculty
14    14  50   male   staff
15    15  60   male faculty
```

#The following command shows male data only

```
> sexm=subset(empinfo,empinfo$sex=='male')
> sexm    #it shows Male data only
   empid age  sex  status
1      1  30 male   staff
3      3  45 male faculty
7      7  35 male   staff
8      8  32 male faculty
10    10  43 male   staff
11    11  32 male faculty
14    14  50 male   staff
15    15  60 male faculty
```

#The following command shows female data only

```
> sexf=subset(empinfo,empinfo$sex=='female')
> sexf
   empid age    sex  status
2      2  37 female   staff
4      4  32 female faculty
5      5  50 female   staff
6      6  60 female   staff
9      9  34 female faculty
12    12  30 female   staff
13    13  43 female faculty
```

Similarly create staff data set and faculty dataset:

Summary statistics for empinfo data

```
> summary(empinfo)
     empid            age              sex         status
 Min.   : 1.0   Min.   :30.00   male  :8   staff  :8
 1st Qu.: 4.5   1st Qu.:32.00   female:7   faculty:7
 Median : 8.0   Median :37.00
 Mean   : 8.0   Mean   :40.87
 3rd Qu.:11.5   3rd Qu.:47.50
 Max.   :15.0   Max.   :60.00
```

Summary statistics for male and female employees data

```
> summary(sexf)
     empid            age              sex         status
 Min.   : 2.000   Min.   :30.00   male  :0   staff  :4
 1st Qu.: 4.500   1st Qu.:33.00   female:7   faculty:3
 Median : 6.000   Median :37.00
 Mean   : 7.286   Mean   :40.86
 3rd Qu.:10.500   3rd Qu.:46.50
 Max.   :13.000   Max.   :60.00
> summary(sexm)
     empid            age              sex         status
 Min.   : 1.000   Min.   :30.00   male  :8   staff  :4
 1st Qu.: 6.000   1st Qu.:32.00   female:0   faculty:4
 Median : 9.000   Median :39.00
 Mean   : 8.625   Mean   :40.88
 3rd Qu.:11.750   3rd Qu.:46.25
 Max.   :15.000   Max.   :60.00
```

Summary statistics for age

> summary(empinfo$age)

Min. 1st Qu. Median Mean 3rd Qu. Max.
30.00 32.00 37.00 40.87 47.50 60.00

Creating one-way table

1. **For sex**

```
> table1=table(empinfo$sex)
> table1

  male female
     8      7
```

2. **For status**

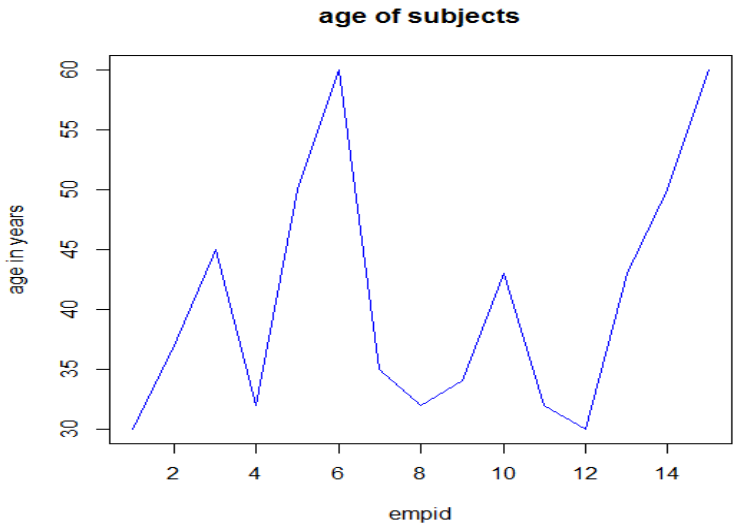
```
> table2=table(empinfo$status)
> table2

  staff faculty
      8       7
```

Creating two-way table

```
> table3=table(empinfo$sex,empinfo$status)
> table3

        staff faculty
male        4       4
female      4       3
```

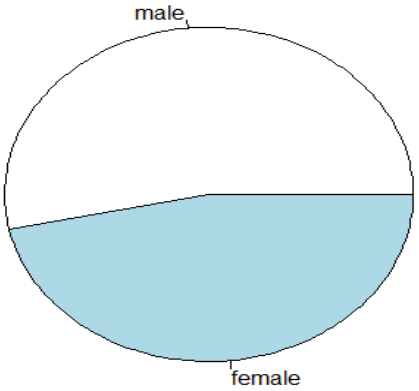
Graphical representation in R:

>plot(empinfo$age,type="l",main="age of subjects",xlab="empid",ylab="age in years",col="blue")



Pie Chart:-

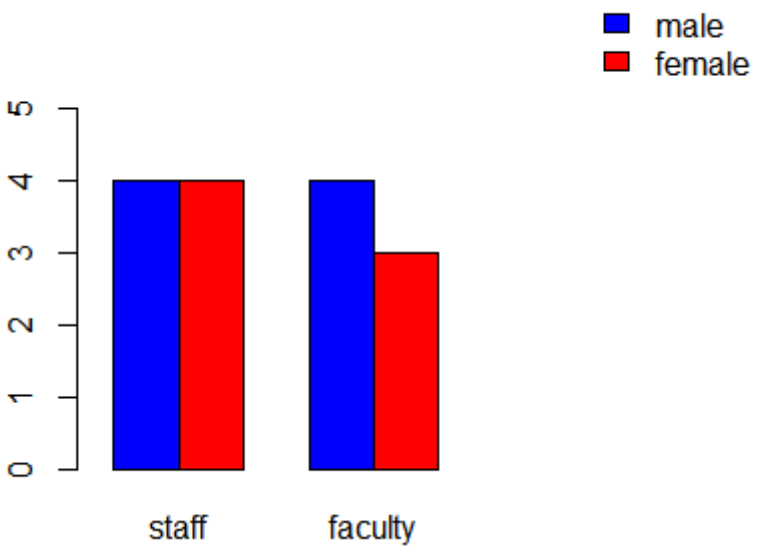
> table4<-table(empinfo$sex)
> pie(table4)



> table5=table(empinfo$sex,empinfo$status)

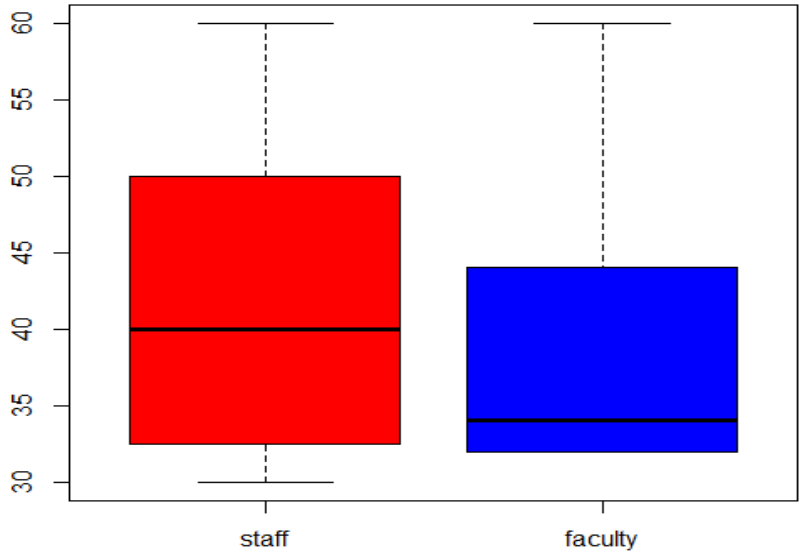
> barplot(table5,beside=T,xlim=c(1,15),ylim=c(0,5),col=c(‘blue’, ‘red’))

>legend("topright",legend=rownames(table5),fill=c('blue','red'),bty="n")



BOXPLOT:-

>boxplot(empinfo$age~empinfo$status,col=c(‘red’,‘blue’))



For practice: Draw Histogram, Frequency polygon for the above data.

Create your own (Student Record) dataset and do the summary statistics and graphs with interpretation. Use at least 50 observations with five variables.

3. Random Variable and Probability Distributions

AIM:

1) Conducting random experiments with probability concepts.
2) Computing and Plotting Binomial and Poisson Distributions.

BASICS IN PROBABILITY:-

1. If you want select five numbers at random from the set 1 to 50 use the command ‘sample’.

   sample(x, size) Example : **>sample(1:50,5)**

2. Sampling with replacement is suitable for modelling coin tosses or throws of a die.

   sample(1:6,10,replace=TRUE) ## replace=FALSE wont work because sample size greather than outcomes. (10 > c(1,2,3,4,5,6))

3. roll 2 dice

   dice=as.vector(outer(1:6,1:6,paste)) ## sample space of rolling two dice
   dice=as.vector(outer(1:6,1:6)) ## product of face values when rolling two dice

4. Toss a coin

   sample(c('H','T'),10,replace=TRUE)

5. Probabilities for the outcomes (chance of success) by using the ‘prob’ argument to sample

   sample(c("success", "fail"), 10, replace=T, prob=c(0.9, 0.1)) ##(replace=F wont work because sample size > outcomes) output with probability of success is 0.9)

   sample(c("success", "fail"), 10, replace=T) ## no restriction on output.

6. Combination for nCr

   choose(n,r)

7. Permutation (no direct command for permutation in R so use ‘factorial’ command)

   For Example: To find 10P5
   n=10;k=5;
   p=factorial(10)/factorial(5)
   p

8. To find the binomial co efficient use ‘choose’ command

   choose(10,0:10) ## for n = 10 and x ranges from 0 to 10.

9. Use choose command to form the Pascal’s triangle with for loop.

   for (n in 0:N) print (choose(n,0:n)) ## N is a positive integer.

10. Tossing ‘n’ coins

    library(prob)
    tosscoin(n)
    [OR]
    prob::tosscoin(n) ## output without probabilities.

11. Tossing ‘n’ coins with probabilities.

    tosscoin(n,makespace=TRUE) ##output with probabilities

12. Roll ‘n’ dice

    rolldie(n) ## for n = 1 to 4 (output restricted upto 4 dice only)

13. Roll ‘n’ dice to restrict the size of the matrix

    rolldie(n,nsides=m) ## for n = 1 to 4 and m= 1 to 6

14. Roll ‘n’ dice with equal probabilities

    rolldie(n,makespace=TRUE)

15. To find expectation and variance for Discrete Random Variable.

```
x=c(0,1,2,3)
p=c(1/8,3/8,3/8,1/8)
mean=sum(x*p)
mean
variance=sum((x^2*p))-(mean^2)
variance
```

4. Discrete and Contributions Probability Distributions

Binomial Distribution

The **binomial distribution** is a discrete probability distribution. It describes the outcome of *n* independent trials in an experiment. Each trial is assumed to have only two outcomes, either success or failure. If the probability of a successful trial is *p*, then the probability of having *x* successful outcomes in an experiment of *n* independent trials is as follows.

SYNTAX

$$P[X=x]=\binom{n}{x}p^x q^{n-x}, x=0,1,.....n$$

Mean $\mu_1^{'}=np$ **Variance:** $\mu_2=npq$

For a binomial(n,p) random variable X, the R functions involve the abbreviation "binom":

```
dbinom(k,n,p)  # binomial(n,p) density at k: Pr(X = k)

pbinom(k,n,p)  # binomial(n,p) CDF at k: Pr(X <= k)

qbinom(P,n,p)  # binomial(n,p) P-th quantile

rbinom(N,n,p)  # N binomial(n,p) random variables

help(Binomial) # documentation on the functions related
       # to the Binomial distribution
```

Problem1. Find the Probability of getting two '2' among ten dice
 # Syntax is dbinom and n=10, x=2 p=1/6,

In general, the syntax is

dbinom(x,size=n,prob=p)

>dbinom(2,size=10,prob=1/6)

[1] 0.29071

Problem 2: Find the P(2) by using binomial probability formula

 # To fit in the general formulat $P[X=x]=nC_x p^x q^{n-x}$ where n=10,x=2,p=1/6, In general, syntax is
choose(n,x)*(p)^x*(q)^n-x
Example
> choose(10,2)*(1/6)^2*(5/6)^8

[1] 0.29071

Problem 3:

Find the table for BIN(n=10,P=1/6) # to list the binomial distribution values as a table, General syntax is evaluate binomial using

n=10;p=1/6

probs=dbinom(x=c(0:n),size=n,prob=p) and use dataframe.

probs=round(probs,4); x=0:n;data.frame(x,probs)

Synatax:

probs=dbinom(x= c(0:10),size=10,prob=1/6)

data.frame(x,probs)

VISUALISATION OF BINOMIAL PLOTS

Problem1: Draw a Plot for the Binomial distribution Bin(n=10,p=1/6)

plot(0:10,probs,type="h",xlim=c(0,10),ylim=c(0,0.5))

 # visualizing 0 to 10 with the type h as a histogram

- **points(0:10,probs,pch=16,cex=2) ## each x value is taken as a character and pch points it. cex expands the character(it changes the shape of the pch) (pch can be any filled shape like circle, square, diamond etc…)**

Problem2: Plot Binomial distribution with n=50 and P=0.33

```
x=0:50
y=dbinom(x,size=50,prob=0.33)
plot(x,y,type="h")
```

Problem3: For a Binomial(7,1/4) random variable named X,
i. Compute the probability of two success
ii. Compute the Probablities for whole space
iii. Display those probabilities in a table
iv. Show the shape of this binomial Distribution
Solution:

(i) dbinom(2,7,1/4) # probability of two success

[1] 0.3114624

(ii) p=dbinom(0:7,7,1/4) # probabilities for whole space

[1] 1.334839e-01 3.114624e-01 3.114624e-01 1.730347e-01 5.767822e-02 1.153564e-02 1.281738e-03 6.103516e-05
round(p,4)
(iii) p=data.frame(0:7,dbinom(0:7,7,1/4))

round(p,4)

(iv) plot(0:7,dbinom(0:7,7,1/4),type="o")

Problem4: Suppose there are twelve multiple choice questions in an English class quiz. Each question has five possible answers, and only one of them is correct. Find the probability of having four or less correct answers if a student attempts to answer every question at random.

Solution
Since only one out of five possible answers is correct, the probability of answering a question correctly by random is 1/5=0.2. We can find the probability of having exactly 4 correct answers by random attempts as follows.

dbinom(4, size=12, prob=0.2)
[1] 0.1329

Problem5: find the probability of having four or less correct answers by random attempts, we apply the function dbinom with $x = 0, \ldots, 4$.

dbinom(0, size=12, prob=0.2) +dbinom(1, size=12, prob=0.2) + dbinom(2, size=12, prob=0.2) + dbinom(3, size=12, prob=0.2) + dbinom(4, size=12, prob=0.2)
[1] 0.9274

OR

Alternatively, we can use the cumulative probability function for binomial distribution pbinom.

pbinom(4, size=12, prob=0.2)
[1] 0.92744

The probability of four or less questions answered correctly by random in a twelve question multiple choice quiz is 92.7%.

Problem 6: If 10% of the Screws produced by an automatic machine are defective, find the probability that out of 20 screws selected at random, there are
(i) Exactly 2 defective
(ii) At least 2 defectives
(iii) Between 1 and 3 defectives (inclusive)

Solution:
(i) > # Exactly 2 defective P=0.10 n=20
dbinom(2,20,0.10)
[1] 0.2851798

(ii) P(X>=2)=1-P(X<2)=1-P(X<=1)=1-P(0:1)
P(X<=2)= pbinom(2,20,0.10,lower.tail=T)
P(X>2)=pbinom(2,20,0.1,lower.tail=F)
1-sum(dbinom(0:1,20,0.1)) (OR)
1-pbinom(1,20,0.1)
[1] 0.608253

(iii) #Between 1 and 3 defectives (inclusive)

x=sum(dbinom(1:3,20,0.10))
x
[1] 0.74547

Relationship between mean and variance :-

Show that Binomial distribution variance is less than mean with Binomial variable follows (7,1/4)

Solution:

```
n=7
p=1/4
px=dbinom(0:n,n,p)
x=0:n
Ex=sum(x*px)
Ex
var=sum(x^2*px)-Ex^2
var
```

THE POISON DISTRIBUTION:

If the number of Bernoulli trials of a random experiment is fairly large and the probability of success is small it becomes increasingly difficult to compute the binomial probabilities. For values of n and p such that $n \geq 150$ and $p \leq 0.05$, the poisson distribution serves as an excellent approximation to the binominal distribution.

The random variable X is said to follow the Poisson distribution if and only if

$$p[X = x] = \frac{e^{-\lambda}\lambda^{x}}{\lfloor x}, x = 0,1,2,.....$$

Assumptions:-

1. Number of Bernoulli trials (n) is indefinitely large, $(n \to \infty)$
2. The trials are independent.
3. Probability of success (p) is very small, $(p \to 0)$

   λ = np is constant, $\lambda = np \Rightarrow p = \frac{\lambda}{n}$
4. Mean and variance in poison distribution are equal

Syntax:-

```
dpois(x, lambda, log = FALSE)
ppois(q, lambda, lower.tail = TRUE, log.p = FALSE)
qpois(p, lambda, lower.tail = TRUE, log.p = FALSE)
rpois(n, lambda)
```

Problem1: a. #P(x=5) with parameter 7

p5=dpois(x=5,lambda=7)
[1] 0.1277167

round(p5,4)

b. P(x=0)+P(x=1)+…….+P(x=5)

p5=dpois(x=0:5,lambda=7)

[1] 0.000911882 0.006383174 0.022341108 0.052129252 0.091226192 0.127716668
round(p5,4)

c. P(x<=5)
sum(dpois(0:5,lambda=7))
[1] 0.3007083
Or
round(ppois(q=5,lambda=7,lower.tail=T),4)
[1] 0.3007

d. P(X>5)
ppois(q=5,lambda=7,lower.tail=F)

Problem2 : Check the relationship between mean and variance in Poisson distribution (4) with **n=100**
X.val=0:100
P.val=dpois(X.val,4)
EX=sum(X.val*P.val) #mean
EX
[1] 4
Var=sum((X.val-EX)^2*P.val) #variance
[1] 4
Problem3 : Compute Probabilities and cumulative probabilities of the values between 0 and 10 for the parameter 2 in poisson distribution.

dpois(0:10,2) # probabilities use round command for restricting the
decimal places.
[1] 1.353353e-01 2.706706e-01 2.706706e-01 1.804470e-01 9.022352e-02
[6] 3.608941e-02 1.202980e-02 3.437087e-03 8.592716e-04 1.909493e-04
[11] 3.818985e-05
(or)
p=data.frame (0:10,dpois(0:10,2))
round (P,4)
ppois(0:10,2) # cumulative probabilities

Problem4: If there are twelve cars crossing a bridge per minute on average, find the probability of having seventeen or more cars crossing the bridge in a particular minute.
Solution:-
The probability of having *sixteen or less* cars crossing the bridge in a particular minute is given by the function ppois.
ppois(16, lambda=12) # lower tail
[1] 0.898709
Hence the probability of having seventeen or more cars crossing the bridge in a minute is in the *upper tail* of the probability density function.
ppois(16, lambda=12, lower=FALSE) # upper tail
[1] 0.101291

Inference: If there are twelve cars crossing a bridge per minute on average, the probability of having seventeen or more cars crossing the bridge in a particular minute is 10.1%.

Problem5: Poisson distribution with parameter 2
1. How to obtain a sequence from 0 to 10
2. Calculate P(0),P(1),...,P(10) when lambda =2 and Make the output prettier
3. Find P(X <= 6)
4. Sum all probabilities
5. Find P(X>6)
6. Make a table of the first 11 Poisson probabilities and cumulative probilities when lamda=2 and obtain the output prettier.
7. Plot the probabilities Put some labels on the axes and give the plot a title

Solution:

1. 0:10 #sequence from 0:10
[1] 0 1 2 3 4 5 6 7 8 9 10

2. round(dpois(0:10, 2), 3)
[1] 0.135 0.271 0.271 0.180 0.090 0.036 0.012 0.003 0.001 0.000 0.000

3. ppois(6, 2) # Find P(x <= 6)
[1] 0.9954662

4. sum(dpois(0:6, 2)) # Sum all probabilities
[1] 0.9954662

5. 1 - ppois(6, 2) # Find P(X>6)
[1] 0.004533806

6. round(cbind(0:10,dpois(0:10,2),ppois(0:10,2)),3)

7. plot(0:10,dpois(0:10,2),type="h",xlab="y",ylab="p(y)",main="Poisson Distribution (mu=2)")

For another type of visualization
points(0:10,dpois(0:10,2),pch=16,cex=2)

Problem6: # TO COMPARE BINOMIAL AND POISSON, USE SAME EXPECTED VALUE.
n=8,
lambda = pn =2.4
# PLOT THE BINOMIAL PMF AND CDF FOR n=8 AND p=0.3
x <- 0:8
px <- dbinom(x,8,0.3); px

[1] 0.05764801 0.19765032 0.29647548 0.25412184 0.13613670 0.04667544 0.01000188
0.00122472 0.00006561

pbinom(x,8,0.3)

[1] 0.05764801 0.25529833 0.55177381 0.80589565 0.94203235 0.98870779 0.99870967
0.99993439 1.00000000

plot(x,px,type="h",col=2,main="Pmf for Binomial(n=8,p=0.3)",xlab="x",ylab="p(x)");
points(x,px,col=2); abline(h=0,col=3)

THE NORMAL DISTRIBUTION :

A random variable X is said to posses normal distribution with mean μ and variance σ^2 , if its probability density function can be expressed of the form,

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} , \quad -\infty < x < +\infty$$

The standard notation used to denote a random variable to follow normal distribution with appropriate mean and variance is, $X \sim N(\mu , \sigma^2)$

STANDARD NORMAL DISTRIBUTION :

If a random variable X follows normal distribution with mean μ and variance σ^2 , its transformation $Z = \frac{X-\mu}{\sigma^2}$ follows standard normal distribution (mean 0 and unit variance)

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

The distribution function of the standard normal distribution

$$F(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx$$

R Synatax :-
R has four in built functions to generate normal distribution. They are described below.

```
dnorm(x, mean, sd)
pnorm(x, mean, sd)
qnorm(p, mean, sd)
rnorm(n, mean, sd)
```

The Normal Distribution

Description
Density, distribution function, quantile function and random generation for the normal distribution with mean equal tomean and standard deviation equal to sd.

Usage
dnorm(x, mean = 0, sd = 1, log = FALSE)
pnorm(q, mean = 0, sd = 1, lower.tail = TRUE, log.p = FALSE)
qnorm(p, mean = 0, sd = 1, lower.tail = TRUE, log.p = FALSE)
rnorm(n, mean = 0, sd = 1)

Arguments

x, q vector of quantiles.

p vector of probabilities.

n number of observations. If length(n) > 1, the length is taken to be the number required.

mean vector of means.

sd vector of standard deviations.

log, log.p logical; if TRUE, probabilities p are given as log(p).

lower.tail logical; if TRUE (default), probabilities are $P[X \leq x]$ otherwise, $P[X > x]$.

Details

If mean or sd are not specified they assume the default values of 0 and 1, respectively.

The normal distribution has density

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

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

Following is the description of the parameters used in above functions:

- ***x*** *is a vector of numbers.*
- ***p*** *is a vector of probabilities.*
- ***n*** *is number of observations(sample size).*
- ***mean*** *is the mean value of the sample data. It's default value is zero.*
- ***sd*** *is the standard deviation. It's default value is 1*

(I) Normal distribution computations and graphs

dnorm() :

This function gives height of the probability distribution at each point for a given mean and standard deviation.

Problem1: Create a sequence of numbers between -10 and 10 incrementing by 0.1

x=seq(-n,n,by=k) # increment is k. Lower and upper values of the sequence are given as – n and n

Example: x= seq(-10,10,by=0.1)

Problem2: Let the mean be 2.5 and standard deviation is 0.5 visualise the normal curve for the above sequence

y=dnorm(x,mean=μ,sd=sigma)

Example:

y=dnorm(x,mean=2.5,sd=0.5)
plot(x,y) # command used to plot the values of x and y

pnorm():

This function gives the probability of a normally distributed random number to be less that the value of a given number. It is also called "Cumulative Distribution Function".

y=pnorm(x,mean=μ,sd=sigma)

x= seq(-10,10,by=0.1)
y=pnorm(x,mean=2.5,sd=0.5)
plot(x,y)

qnorm():-

This function takes the probability value and gives a number whose cumulative value matches the probability value.

y=qnorm(x,mean=μ,sd=sigma)

x=seq(0,1,by=0.02)

y=qnorm(x,mean=2,sd=1)

plot(x,y)

rnorm()

This function is used to generate random numbers whose distribution is normal. It takes the sample size as input and generates that many random numbers. We draw a histogram to show the distribution of the generated numbers.

To create a sample of 50 numbers which are normally distributed

y=rnorm(n) # n is the size or number of observations

hist(y,main="Title-Normal distribution") # hist denotes histogram and main is the title of the plot)

y=rnorm(50)
hist(y,main="Normal distribution")

(II) Standard Normal Probability Distribution Plotting and Finding the Area:

To create a sequence of n numbers with x=-m to x=m with given mean and sd.

x=seq(-m,m,length=n) # length denotes how many numbers we want to create

y=dnorm(x,mean=μ,sd=sigma)

use dnorm to compute the y values of the standard normal and caluculate the pdf with the given mean and sd

plot(x,y)

plot(x,y,type="l") # l means a line graph. Please change it as "h" and check the nature of the curve

Problem1: To create a sequence of 200 numbers with x=-3 to 3 with mean 0 and sd=1

x=seq(-3,3,length=200)
y=dnorm(x,mean=0,sd=1)
plot(x,y)
plot(x,y,type="l")
Output: Bell shaped curve will be created

Problem2: To find the area under the curve to left of the mean

x=seq(-3,3,length=200)
y=dnorm(x,mean=0 ,sd=1)
plot(x,y,type= "l")
x=seq(-3,0,length=100)
y=dnorm(x,mean=0,sd=1)
polygon(c(-3,x,0),c(0,y,0),col="red")
pnorm(0,mean= 0 ,sd=1)

Problem2: To find the area to the left of x=1.
x=seq(-3,3,length=200)
y=dnorm(x,mean=0 ,sd=1)
plot(x,y,type= "l")
x=seq(-3,1,length=100)
y=dnorm(x,mean=0,sd=1)
polygon(c(-3,x,1),c(0,y,0),col="blue")
pnorm(1,mean= 0 ,sd=1)

Problem3: Find the area to the right of x=2. (First draw the image then compute)
x=seq(-3,3,length=200)
y=dnorm(x,mean=0 ,sd=1)
plot(x,y,type= "l")
x=seq(2,3,length=100)
y=dnorm(x,mean=0,sd=1)
polygon(c(2,x,3),c(0,y,0),col="blue")
1-pnorm(2,mean=0,sd=1)
1-pnorm give the area of the right extreme portion of the normal curve with x=2)

FOR PRACTICE
Use pnorm command and find the areas under the curve for any mean and sd

Problem4: To find the Quantile(percentile)- reverse the process which means given the area find the value of x

x=seq(-3,3,length=200)
y=dnorm(x,mean=0 ,sd=1)
plot(x,y,type= "l")
x=seq(-3,-0.2533,length=100)
y=dnorm(x,mean=0,sd=1)
polygon(c(-3,x,-0.2533),c(0,y,0),col="blue")
text(-1,0.2,"0.40") # on x=-1 y=0.2 and 40%)
qnorm(0.40,mean=0,sd=1) # it will calculate the value.

Problem6: Find P(30<X<70) with mean=50 and standard deviation=10.
x=seq(20,80,length=200)
y=dnorm(x,mean=50,sd=10)
plot(x,y,type="l")
x=seq(30,70,length=100)
y=dnorm(x,mean=50,sd=10)
polygon(c(30,x,70),c(0,y,0),col="red")
pnorm(70,mean=50,sd=10)-pnorm(30,mean=50,sd=10)

Problem7: 1. If Z is norm(mean = 0; sd = 1), find
(i) P(Z > 2:64) (ii) P(0 < Z < 0:87) (iii) P(Z> 1:39)
Solution: (i)
pnorm(2.64, lower.tail = FALSE)
0.004145301
(ii) pnorm(0.87) - 1/2
0.3078498
(iii) **2 * pnorm(-1.39)**
0.1645289

Problem8: Find P(Z< -1.24)
Solution:
pnorm(-1.24)
[1] 0.1074877

Draw Graph and shade for the above problem is(WITHOUT USING PLOT COMMAND)
z=-1.24
x = c(-3,seq(-3,z,by=.001),z)
y = c(0, dnorm(seq(-3, z, by=.001)), 0)
polygon(x, y, col="red")

NORMAL PROBABILITY SHAPE:

```
plot.new() # Gives the curve in new figure window
curve(dnorm, xlim = c(-3, 3), ylim = c(0, 0.5), xlab = "z", ylab="f(z)")
 zleft = 0
 zright = 1.24
 x = c(zleft, seq(zleft, zright, by=.001), zright)
 y = c(0, dnorm(seq(zleft, zright, by=.001)), 0)
 polygon(x, y, col="red")
```

Problems:

1. In a photographic process the developing times of prints may be looked upon as a random variable having the normal distribution with a mean of 16.28 seconds and a standard deviation 0.12 second. Find the probability that it will take
(i) at least 16.20 seconds to develop one of the prints;
(ii) at most 16.35 seconds to develop one of the prints

Exercise:

Practice Problems:-(Binomial distribution)

1. For a random variable X with a binomial(20,1/2) distribution, find the following probabilities.
   (i). Find Pr(X < 8)
   (ii). Find Pr(X > 12)
   (iii) Find Pr(8 <= X <= 12)

2. For a binomial(200,1/2) distribution:
   (i) Find Pr(X < 80)
   (ii) Find Pr(X > 120)
   (iii) Find Pr(80 <= X <= 120)

3. For a binomial(2000,1/2) distribution:
   Find Pr(X < 800)
   Find Pr(X > 1200)
   Find Pr(800 <= X <= 1200)

4. Let X be the number of heads in 10 tosses of a fair coin.
   1. Find the probability of getting at least 5 heads (that is, 5 or more).
   2. Find the probability of getting exactly 5 heads.
   3. Find the probability of getting between 4 and 6 heads, inclusive

5. Suppose our random variable X is Poisson with lambda = 12.33.
   1. What is the probability of 15 or fewer occurrences? P(X <= 15)
   2. What is the probability of EXACTLY 6 occurrences? P(X = 6)
   3. What is the probability of more than 15 occurrences? P(X > 15)
   4. What is the probability of 15 or more occurrences? P(X >= 15)
   5. What is the probability of 8, 9, or 10 occurrences? P(8 <= X <= 10)

Compare binomial distribution and Poisson distribution

6. Let X be the number of heads in 100 tosses of a fair coin.
7. Let X be the number of heads in 1000 tosses of a fair coin.

1. *Find (i)* $P(0.8 \le Z \le 1.5)$ *(ii)* $P(Z \le 2)$ *(iii)* $P(Z \ge 1)$
   Find These probability values and Plot the graph .

2. *If mean=70 and Standard deviation is 16*

   i) $P(38 \le X \le 46)$ *ii)* $P(82 \le X \le 94)$ *iii)* $P(62 \le X \le 86)$

   Find the Probability values and Plot the graph with text.

3. *1000 students had Written an examination the mean of test is 35 and standard deviation is 5.Assumning the to be normal find*
   i) How many students Marks Lie between 25 and 40
   ii) How many students get more than 40
   iii) How many students get below 20
   iv) How many students get 50

5. Correlation and Regression

Correlation Definition:-

Correlation refers to the relationship between two or more variables. Simple correlation studies the relationship between two variables. Correlation analysis attempts to determine the degree of relationship between variables.

Measures of Correlation:

Scatter Diagram:

Scatter diagram is the simplest way of graphic representation of a bivariate data, where the given set of 'n' pairs of observations on two variables X and Y say (X_1, Y_2), (X_2, Y_2) ...(X_n, Y_n) may be plotted as dots by considering X-values on X-axis and Y-values on Y-axis. By scatter diagram, we can get some idea about the correlation between X and Y.

Correlation, Variance and Covariance (Matrices)

Description

var, cov and cor compute the variance of x and the covariance or correlation of x and y if these are vectors. If x and y are matrices then the covariances (or correlations) between the columns of x and the columns of y are computed.

cov2cor scales a covariance matrix into the corresponding correlation matrix*efficiently*.

Usage

```
var(x, y = NULL, na.rm = FALSE, use)

cov(x, y = NULL, use = "everything",
    method = c("pearson", "kendall", "spearman"))

cor(x, y = NULL, use = "everything",
    method = c("pearson", "kendall", "spearman"))

cov2cor(V)
```

Arguments

x a numeric vector, matrix or data frame.

y NULL (default) or a vector, matrix or data frame with compatible dimensions to x. The default is equivalent to y = x (but more efficient).

na.rm logical. Should missing values be removed?

use an optional character string giving a method for computing covariances in the presence of missing values. This must be (an abbreviation of) one of the strings "everything", "all.obs", "complete.obs","na.or.complete", or "pairwise.complete.obs".

method a character string indicating which correlation coefficient (or covariance) is to be computed. One of "pearson" (default),"kendall", or "spearman": can be abbreviated.

V symmetric numeric matrix, usually positive definite such as a covariance matrix.

Problem:

Age Group	Representative Age	Hours Spend in Liabrary
10-19	15	302.38
20-29	25	193.63
30-39	35	185.46
40-49	45	198.49
50-59	55	224.30
60-69	65	288.71

Problem1: Plot age (x) and number of hours spent in the local library(y)

x <- c(15,25,35,45,55,65) # input the x values

y <- c(302.38, 193.63, 185.46, 198.49, 224.30, 288.71) #input the y values

>plot(x,y, main="Average age vs. time spent in the library", xlab="Age", ylab="time spent in the library",col="red")

Karl Pearson's Coefficient of Correlation

It is defined as the ratio of covariance between x and y say Cov (X,Y) to the product of the standard deviations of X and Y, say $\sigma_X \sigma_Y$

$$i.e \qquad r_{XY} = \frac{Cov(XY)}{\sigma_X \sigma_Y}$$

Problem2: Find the correlation coefficient for the following data.

X	23	27	28	28	29	30	31	33	35	36
Y	18	20	22	27	21	29	27	29	28	29

Solution:

x=c(23,27,28,28,29,30,31,33,35,36)

y=c(18,20,22,27,21,29,27,29,28,29)

var(x)

15.33333

var(y) # COMMAND FOR FINDING VARIANCE FOR Y

18.22222

var(x,y) # COMMAND FOR FINDING VARIANCE FOR X AND Y

13.66667

r=var(x,y)/sqrt(var(x)*var(y)) # CORRELATION BETWEEN X AND Y

r

[1] 0.8176052

(or)

r=cor(x,y) # CORRELATION BETWEEN X AND Y

r

0.8176052

(or)

cor.test(x,y)

(or)

cor.test(x,y,method="pearson")

SPEARMAN'S RANK CORRELATION COEFFICIENT

Suppose we associate the ranks to individuals or items in two series based on order of merit, the Spearman's Rank correlation coefficient ρ is given by

$$\rho = 1 - \left[\frac{6\sum d^2}{n(n^2-1)}\right]$$ [Read the symbol (as 'Rho'.]

Where, $\sum d^2$ = Sum of squares of differences of ranks between paired items in two series n = Number of paired items'

Problem1: Twelve recruits were subjected to selection test to ascertain their suitability for a certain course of training. At the end of training they were given a proficiency test. The marks scored by the recruits are recorded below.

Recruit	1	2	3	4	5	6	7	8	9	10	11	12
Selection Test Score	44	49	52	54	47	76	65	60	63	58	50	67
Proficiency Test Score	48	55	45	60	43	80	58	50	77	46	47	65

Calculate rank correlation coefficient.

Solution:

selection =c(44,49,52,54,47,76,65,60,63,58,50,67)

proficiency =c(48,55,45,60,43,80,58,50,77,46,47,65)

cor.test(selection,proficiency,method="spearman")

Regression:

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.

Solution:

```
weight=c(15,26,27,2,25.5,27,32,18,22,20,26,24)
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)
coe(weight,bmi)
mode1<-lm(bmi~weight)  # lm means linear model
summary.lm(mode1)
```

OUTPUT

```
Call:
lm(formula = bmi ~ weight)

Residuals:
    Min      1Q  Median      3Q     Max
-33.838 -10.253  -6.582  -2.659  99.734

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   52.334     30.978   1.689    0.122
weight        -1.248      1.331  -0.937    0.371

Residual standard error: 34.39 on 10 degrees of freedom
Multiple R-squared:  0.08076,           Adjusted R-squared:  -0.01116
F-statistic: 0.8786 on 1 and 10 DF,  p-value: 0.3707
```

Interpretation :

Correlation r=0.5790,which is the correlation coefficient between the body 'weight' and BMI.There is a positive correlation between these two variables.The Value of R^2 is 0.3353,which means that about 33.53% variation in BMI can be explained by 'weight'through this linear model.This is apparently low because more than 67% of variation remains unexplained.There could be several reasons for this and one of them is that there might be some other influencing variables that have not been included in the present model.

The F value shown in the above output gives the statistics for the variance ratio test of the regression model.The significance of F, which is given as 0.0485,is the p value of the F-test carried out in ANOVA. If this value is less than 0.05 we say that the regression is statistical significant at 5% level of significance .Here

regression is significant which means that the relationship is not an occurrence by chance

In the above output we find b_0 is the intercept which value of 10.73487 and b_1 is the regression coefficient due to weight with a value of 0.1710.The regression coefficient is positive ,which shows that the BMI is positively related to weight,

The regression output can be written as mathematical equation

$$BMI = 10.7349 + 0.1710 * weight$$

Suppose body weight of one student is known as 25 kg.Using the above equation,the estimated BMI is 15.01.since this is only an estimate we have to intercept it as the average BMI corresponding to the given weight assuming that other parameters are unchanged.

Exercise:

1. ***The following data refers to the daily sales of tomatoes (in kg) at different prices(in Rupess) observed on different days in a market***

Price	*4.5*	*5.5*	*4.5*	*4.5*	*4.0*	*5.5*	*5.5*	*6.5*	*5.0*	*5.5*	*6.0*	*4.5*
Quantity Sold	*125*	*115*	*140*	*140*	*150*	*150*	*130*	*120*	*130*	*100*	*105*	*150*

2 Obtain a linear relationship between weight (kg) and height (cm) of 10 subjects.

Height	*175*	*168*	*170*	*171*	*169*	*165*	*165*	*160*	*180*	*186*
Weight	*80*	*68*	*72*	*75*	*70*	*65*	*62*	*60*	*85*	*90*

6. Multiple Linear Regression

Description

lm is used to fit linear models. It can be used to carry out regression, single stratum analysis of variance and analysis of covariance (although aov may provide a more convenient interface for these).

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

Solution:

Y=c(110,80,70,120,150,90,70,120)

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

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

input_data=data.frame(Y,X1,X2)

input_data

RegModel<- lm(Y~X1+X2, data=input_data)

RegModel

(OR) use this statement:

lm(formula = Y ~ X1 + X2, data = input_data)

summary(RegModel)

Note: It will display the Residual standard error,Multiple R-squared value,Ajusted R-squared value,F-statistic values

Interpretation :

Now the regression the regression model is

$$Y = 16.834 - 0.2442 * X1 + 7.8488 * X2$$

Since R^2 is 0.9593 and the ANOVA shows that the F-ratio is significant, this model can be taken as good-fit in explaining the sales interms of the other two variables

Problem 2 :(Health.csv): Let us develop a multiple regression model of BMR on the variables age, HT, WT and BMI and interpret the data. (Health.csv file has to be given)

Solution:

data=read.csv(file.choose()) or data=read.csv("c:/Users/admin/Desktop/health.csv")

regmodel=lm(BMR~AGE+HT+WT+BMI,data=data)

regmodel

summary(regmodel)

Interpretation:

Now the Regression model can be stated as

$$BMR = -2500.492 + 4.021(age) + 17.293(HT) + 1.1019 + 50.553(BMI)$$

R^2 is 0.8701 ,which is about 87% of BMR can be explained in terms of age HT,WT and BMI of a person through this linear model,we also see that all the explanatory variables have positive relationship with BMR.These regression coefficient are how ever not statistically significant except that of age,though the F-test in ANOVA shows that the overall regression is significant at 0.01 level(p-value is almost zero).The meaning of the regression coefficient can be understood as follows

if the age increases by 4.021 at fixed values of the other factors like HT,WT and BMI.

7. Testing of Hypothesis- I (Z test)

Test for significance of single mean:

Lower Tail Test of Population Mean with Known Variance:

Null Hypothesis: $\mu \geq \mu_0$.

Alternative hypothesis*:* $\mu < \mu_0$. μ_0 is a hypothesized lower bound of the true population mean μ.

Test Statistic: $Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$. Reject Null hypothesis if $z \leq -z_\alpha$.

Problem

Suppose the manufacturer claims that the mean lifetime of a light bulb is more than 10,000 hours. In a sample of 30 light bulbs, it was found that they only last 9,900 hours on average. Assume the

population standard deviation is 120 hours. At 0.05 significance level, can we reject the claim by the manufacturer?

Null hypothesis*:* $\mu \geq 10000$ .Alternative Hypothesis: $\mu < 10000$

R-code

```
xbar=9900
mu0=10000
sigma=120
n=30
z=(xbar-mu0)/(sigma/sqrt(n))
z
[1] -4.564355
alpha=0.05
zalpha=qnorm(1-alpha)
-zalpha
[1] -1.644854
pval=pnorm(z)
pval
[1] 2.505166e-06
```

Interpretation: The test statistic -4.5644 is less than the critical value of -1.6449. Hence, at 0.05 significance level, we reject the claims that mean lifetime of a light bulb is above 10,000 hours.

The lower tail **p-value** of the test statistic is less than the significance level 0.05, we reject the null hypothesis that $\mu \geq 10000$.

Upper Tail Test of Population Mean with Known Variance:

Null hypothesis*:* $\mu \leq \mu_0$. μ_0 is a hypothesized upper bound of the true population mean μ.

Test Statistic: $Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$ N(0,1) . Reject Null hypothesis if $z \geq z_\alpha$.

Problem

Suppose the food label on a cookie bag states that there is at most 2 grams of saturated fat in a single cookie. In a sample of 35 cookies, it is found that the mean amount of saturated fat per cookie is 2.1 grams. Assume that the population standard deviation is 0.25 grams. At 0.05 significance level, can we reject the claim on food label?

R-code

```
xbar=2.1
mu0=2
sigma=0.25
n=35
z=(xbar-mu0)/(sigma/sqrt(n))
 z
[1] 2.366432
alpha=0.05
zalpha=qnorm(1-alpha)
```

zalpha
[1] 1.644854
pval=pnorm(z)
pval
[1] 0.9910198
1-pval
[1] 0.008980239

Interpretation:-
The test statistic 2.3664 is greater than the critical value of 1.6449. Hence, at 0.05 significance level, we reject the claim that there is at most 2 grams of saturated fat in a cookie.

The upper tail **p-value** of the test statistic is less than the significance level 0.05, we reject the null hypothesis that $\mu \leq 2$.

Two-Tailed Test of Population Mean with Known Variance:-

The null hypothesis of the **two-tailed test of the population mean** can be expressed as follows: $\mu = \mu_0$ .where $\mu = \mu_0$ is a hypothesized value of the true population mean μ.

Test Statistic: $Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$ □ N(0,1) . Reject Null hypothesis if $z \leq -z_{\alpha/2}$ or $z \geq z_{\alpha/2}$.

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 .05 significance level, can we reject the null hypothesis that the mean penguin weight does not differ from last year?

xbar=14.6
mu0=15.4
sigma=2.5
n=35
z=(xbar-mu0)/(sigma/sqrt(n))
z

[1] -1.893146

alpha=0.05
zhalfalpha=qnorm(1-(alpha/2))
c(-zhalfalpha,zhalfalpha)

[1] -1.959964 1.959964

pval=2*pnorm(z)
pval

[1] 0.05833852

Interpretation :

The test statistic -1.8931 lies between the critical values -1.9600 and 1.9600. Hence, at 0.05 significance level, we do not reject the null hypothesis that the mean penguin weight does not differ from last year.

p-value is greater than the 0.05 significance level, we do not reject the null hypothesis that μ = 15.4.

Lower Tail Test of Population Proportion:

The null hypothesis of the **lower tail test about population proportion** can be expressed as follows: $p \geq p_0$.where p_0 is a hypothesized lower bound of the true population proportion p.

Test Statistic: $Z = \frac{p - p_0}{\sqrt{p_0 q_0 / n}} \sim N(0,1)$. Reject Null hypothesis if $z \leq -z_\alpha$.

Problem
Suppose 60% of citizens voted in last election. 85 out of 148 people in a telephone survey said that they voted in current election. At 0.05 significance level, can we reject the null hypothesis that the proportion of voters in the population is above 60% this year?

p=85/148
p0=60/100
n=148
q0=1-p0
z=(p-p0)/sqrt(p0*q0/n)
z
[1] -0.6375983
alpha=0.05
zalpha=qnorm(1-alpha)
-zalpha
[1] -1.644854
pval=pnorm(z)
pval
[1] 0.2618676

Interpretation :
The test statistic -0.6376 is not less than the critical value of -1.6449. Hence, at 0.05 significance level, we do not reject the null hypothesis that the proportion of voters in the population is above 60% this year.

p-value is greater than the 0.05 significance level, we do not reject the null hypothesis that the proportion of voters in the population is above 60% this year.

Upper Tail Test of Population Proportion

The null hypothesis of the **upper tail test about population proportion** can be expressed as follows: as follows: $p \leq p_0$.where p_0 is a hypothesized upper bound of the true population proportion p.

Test Statistic: $Z = \frac{p - p_0}{\sqrt{p_0 q_0 / n}} \sim N(0,1)$. Reject Null hypothesis if $z \geq z_\alpha$.

Problem
Suppose that 12% of apples harvested in an orchard last year was rotten. 30 out of 214 apples in a harvest sample this year turns out to be rotten. At .05 significance level, can we reject the null hypothesis that the proportion of rotten apples in harvest stays below 12% this year?

The null hypothesis is that $p \leq 0.12$.

p=30/214
p0=12/100
q0=1-p0
n=214
z=(p-p0)/sqrt(p0*q0/n)
z
[1] 0.908751
alpha=0.05
zalpha=qnorm(1-alpha)
zalpha
[1] 1.644854
pval=pnorm(z,lower.tail=FALSE)
pval
[1] 0.1817408

Interpretation:-
The test statistic 0.90875 is not greater than the critical value of 1.6449. Hence, at 0.05 significance level, we do not reject the null hypothesis that the proportion of rotten apples in harvest stays below 12% this year.

p-value is greater than the 0.05 significance level, we do not reject the null hypothesis that the proportion of rotten apples in harvest stays below 12% this year.

Two-Tailed Test of Population Proportion:

The null hypothesis of the **two-tailed test about population proportion** can be expressed as follows: $p = p_0$

Test Statistic: $Z = \frac{p - p_0}{\sqrt{p_0 q_0 / n}} \sim N(0,1)$. Reject Null hypothesis if $z \leq -z_{\alpha/2}$ or $z \geq z_{\alpha/2}$.

Problem
Suppose a coin toss turns up 12 heads out of 30 trials. At 0.05 significance level, can one reject the null hypothesis that the coin toss is fair?

The null hypothesis is that p = 0.5.

p=18/30
p0=1/2
q0=1-p0
z=(p-p0)/sqrt(p0*q0/n)
z
[1] 1.095445
alpha=0.05
zhalfalpha=qnorm(1-(alpha/2))

c(-zhalfalpha,zhalfalpha)
[1] -1.959964 1.959964
pval=2*pnorm(z,lower.tail=FALSE)
pval
[1] 0.2733217

Interpretation:
The test statistic 1.095445 lies between the critical values -1.9600 and 1.9600. Hence, at 0.05 significance level, we do not reject the null hypothesis that the coin toss is fair.

p-value is greater than the 0.05 significance level, we do not reject the null hypothesis that the coin toss is fair.

8. Testing of Hypothesis – II(student‘t, F, Chi-square)

AIM: To test the single mean, difference of means for small sample using t-test and also test the hypothesis for variance ratio using F-test.

Problem1: An outbreak of salmonella-related illness was attributed to ice produced at a certain factory. Scientists measured the level of Salmonella in 9 randomly sampled batches ice crean.The levels(in MPN/g) were:

0.593	0.142	0.329	0.691	0.231	0.793	0.519	0.392	0.418

Is there evidence that the mean level pf Salmonella in ice cream greater than 0.3 MPN/g?

Solution:

H_0: mu=0.3, H_1:mu>0.3

```
x=c(0.593,0.142,0.329,0.691,0.231,0.793,0.519,0.392,0.418)
xbar=mean(x)
alpha=0.05
mu=0.3
sd=sqrt(var(x))
n=length(x)
t=(xbar-mu)/(sd/sqrt(n))
t
tv=qt(1-alpha,df=n-1)
tv
```

Inference:-

From the output we see that the t = 2.205059>1.859548. Reject H_0.Hence, there is moderately strong evidence that the mean Salmonella level in the ice cream is above 0.3MPN/g.

Problem2: Suppose that 10 volunteers have taken an intelligence test; here are the results obtained. The average score of the entire population is 75 in the same test. Is there any significant difference (with a significance level of 95%) between the sample and population means, assuming that the variance of the population is not known.

Scores: 65, 78, 88, 55, 48, 95, 66, 57, 79, 81

Solution:

H_0: mu=75, H_1:mu≠75

```
x=c(65, 78, 88, 55, 48, 95, 66, 57, 79, 81)
xbar=mean(x)
sd=sqrt(var(x))
mu=75
```

```
alpha=0.05
n=length(x)
t=abs(xbar-mu)/(sd/sqrt(n))
t
qt(1-(alpha/2), n-1)
```

Inference:-

The t-computed value is smaller than t-tabulated, we accept the null hypothesis of equality of the averages.

Problem3: Comparing two independent sample means, taken from two populations with unknown variance. The following data shows the heights of individuals of two different countries with unknown population variances. Is there any significant difference b/n the average heights of two groups.

A:	175	168	168	190	156	181	182	175	174	179
B:	185	169	173	173	188	186	175	174	179	180

Solution:

$H_0: \mu_1 = \mu_2$, $H_1: \mu_1 \neq \mu_2$

```
x1=c(175,168,168,190,156,181,182,175,174,179)
x2=c(185,169,173,173,188,186,175,174,179,180)
alpha=0.05
n1=length(x1)
n2=length(x2)
x1bar=mean(x1)
x2bar=mean(x2)
sd1=sqrt(var(x1))
sd2=sqrt(var(x2))
t=abs(x1bar-x2bar)/(sqrt((sd1^2/n1)+(sd2^2/n2)))
tv=qt(1-(alpha/2),n1+n2-2)
```

Inference :

The calculated value of t is less than the tabulated t-value for 18 df, we accept H_0 and conclude that there is no significant difference between the average heights of two groups.

Problem4: Suppose the recovery time for patients taking a new drug is measured (in days). A placebo group is also used to avoid the placebo effect. The data are as follows

with drug	: 15 10 13 7 9 8 21 9 14 8
placebo	: 15 14 12 8 14 7 16 10 15 2

Test whether the recovery time of new drug group is greater than the placebo group?

Solution:

$H_0: \mu_1 = \mu_2$, $H_1: \mu_1 > \mu_2$

```
x=c(15,10,13,7,9,8,21,9,14,8)
y=c(15, 14,12,8,14,7,16,10,15,2)
alpha=0.05
n1=length(x)
n2=length(y)
xbar=mean(x)
ybar=mean(y)
sd1=sqrt(var(x))
sd2=sqrt(var(y))
t=abs(xbar-ybar)/(sqrt((sd1^2/n1)+(sd2^2/n2)))
tv=qt(1-alpha,n1+n2-2)
```

Inference :-

The calculated value of t is less than the tabulated t-value for 18 df, we accept H_0 and conclude that there is no significant difference between the average recovery time of two groups.

F-Test

Problem1 :-

Five Measurements of the output of two units have given the following results (in kilograms of material per one hour of operation) .Assume that both samples have been obtained from normal populations, test at 10% significance level if two populations have the same variance.

Unit A	14.1	10.1	14.7	13.7	14.0
Unit B	14.0	14.5	13.7	12.7	14.1

Solution:

$H_0 : S_1^2 = S_2^2, H_1 : S_1^2 \neq S_2^2$

```
A=c(14.1,10.1,14.7,13.7,14.0)
B=c(14.0,14.5,13.7,12.7,14.1)
alpha=0.01
n1=length(A)
n2=length(B)
TV=qf(1-alpha,n1-1,n2-1)
F=var(A)/var(B)
```

Inference : Here TV >F value ,then there is no evidence to reject the null hypothesis.
Hence, the two samples came from a same population.

Problem2: In order to compare the effectiveness of two sources of nitrogen, namely ammonium chloride and urea on grain yield of paddy, an experiment was conducted. The results on the grain yield of paddy(kg/plot) under the two treatments are given below
Ammonium
Chloride: 13.4 10.9 11.2 11.8 14.0 15.3 14.2 12.6 17.0 16.2 16.5 15.7
Urea : 12.0 11.7 10.7 11.2 14.8 14.4 13.9 13.7 16.9 16.0 15.6 16.0
Asses which sources nitrogen is better for paddy.

Solution:

$H_0 : S_1^2 = S_2^2, H_1 : S_1^2 \neq S_2^2$

```
x=c(13.4,10.9,11.2,11.8,14.0,15.3,14.2,12.6,17.0,16.2,16.5,15.7)
y=c(12.0,11.7,10.7,11.2,14.8,14.4,13.9,13.7,16.9,16.0,15.6,16.0)
alpha=0.05
n1=length(x)
n2=length(y)
TV=qf(1-alpha,n1-1,n2-1)
F=var(A)/var(B)
```

Inference : Here TV >F value ,then there is no evidence to reject the null hypothesis.
Hence, two sources of nitrogen give the same result. Lowest cost nitrogen source can be used.

CHI-SQUARE TEST

(INDEPENDENCE OF ATTRIBUTES AND GOODNESS OF FIT)

Aim: To test is there any association between the attributes (or) Independence of attributes and Goodness of fit using chi-square distribution

$$\chi^2 = \sum\left[\frac{(O_{ij} - E_{ij})^2}{E_{ij}}\right]$$

Problem 1 :The below table gives the distribution of students according to the family type and the anxiety level

Family type	*Anxiety level*		
	Low	*Normal*	*High*
Joint family	*35*	*42*	*61*
Nuclear family	*48*	*51*	*68*

data<-matrix(c(35,42,61,48,51,68),ncol=3,byrow=T)
data
chisq.test(data)

Interpretation :-

Here P value (0.7655) > 0.05. Hence there is no evidence to reject the Null hypothesis. So we consider the anxiety level and family type as independent.

Problem2:

In the built-in data set survey, the Smoke column records the students smoking habit, while the Exer column records their exercise level. The allowed values in Smoke are "Heavy", "Regul" (regularly), "Occas" (occasionally) and "Never". As for Exer, they are "Freq" (frequently), "Some" and "None".We can tally the students smoking habit against the exercise level with the table function in R. The result is called the contingency table of the two variables.

```
library(MASS)
tb1=table(survey$smoke,survey$Exer)
tb1
chisq.test(tb1)
```

Interpretation :-

As the p-value 0.4828 is greater than the .05 significance level, we do not reject the null hypothesis that the smoking habit is independent of the exercise level of the students.

Enhanced Solution:

The warning message found in the solution above is due to the small cell values in the contingency table. To avoid such warning, we combine the second and third columns of tbl.

```
ctb1=cbind(tb1[,"Freq"],tb1[,"None"],tb1[,"Some"])
ctb1
chisq.test(tb1)
```

Problem 3 :

A biologist is conducting a plant breeding experiment in which plants can have one of four phenotypes. If these phenotypes are caused by a simple Mendelian model, the phenotypes should occur in a 9:3:3:1 ratio. She raises 41 plants with the following phenotypes.

Phenotype	1	2	3	4
count	20	10	7	4

Should she worry that the simple genetic model doesn't work for her phenotypes?

```
plants<-c(20,10,7,4)
chisq.test(plants,p=c(9/16,3/16,3/16,1/16))
```

Note: Please ignore the warning message in the console window after executing the program.

Fitting of Binomial distribution with goodness of fit:-

Problem 4 : *A survey of 320 families with 5 children each revealed the following distribution:*

Number of Boys	5	4	3	2	1	0
No of Girls	0	1	2	3	4	5
No of families	14	56	110	88	40	12

Is this result consistent with the hypothesis that male and female births are equally possible?

```
n=5   # probability of r  male births in a family
alpha=0.05
N=320                 # Total number of families
P<-0.5                # probability of male birth
x=c(0:n)
obf<-c(14,56,110,88,40,12)                  # observed frequencies
exf<-(dbinom(x,n,P)*320)                    # expected frequencies
# check the condition if the observed and expected frequencies sum are equal sum(obf)
sum(exf)
chisq<-sum((obf-exf)^2/exf)
cv=chisq
tv=qchisq(1-alpha,n-1)
if(cv <= tv){print("Accept Ho")} else{print("Reject Ho")}
```

Intrepretation: The binomial distribution is the best fit for the given data.

Fitting of Poisson distribution with goodness of fit:-

Problem 5: Fit a Poisson distribution to the following data and test the goodness of fit

X	0	1	2	3	4	5	6
f	275	72	30	7	5	2	1

```
x<-0:6
f<-c(275,72,30,7,5,2,1)
lambda<-(sum(f*x)/sum(f))          # mean
expf= dpois(x,lambda)*sum(f)     # expected frequencies
f1=round(expf)
sum(f)
# hints : here subtract “1” from the expected frequencies and the last three #frequencies are less than 5 so combine these frequencies in observed and #expected)
obf<-c(275,72,30,15)
exf<-c(242,117,28.6)
```

```
chisq<-sum(((obf-exf)^2)/exf)
cv=chisq
tv=qchisq(0.95,1)
if(cv <= tv){print("Accept Ho")} else{print("Reject Ho")}
```

Inference: The Poisson distribution is not fit for the given data.

9. COMPLETELY RANDOMIZED DESIGN

MODEL FOR ONE WAY ANOVA

AIM:

To find the ANOVA using CRD to test the null hypothesis (H_0) against alternative hypothesis(H_1) with level of significance , alpha=0.05.

MODEL:

$Y_{ij} = \mu + \alpha_i + \in_{ij}$ where $\in_{ij}$ is the deviation of the jth observation of the ith sample from the corresponding treatment mean (ie) random error, Y_{ij} jth observation from the ith treatment, μ is the grand mean and α_i is the effect of ith treatment.

PROCEDURE:

The null hypothesis that the k population means are equal against the alternative that at least two of the means are unequal.

The null hypothesis is rejected if $P = P\{F[k-1, k(n-1)] > F\}$ *or* $\Pr(>F)$ is less than alpha value or else null hypothesis is accepted.

SYNTAX USED

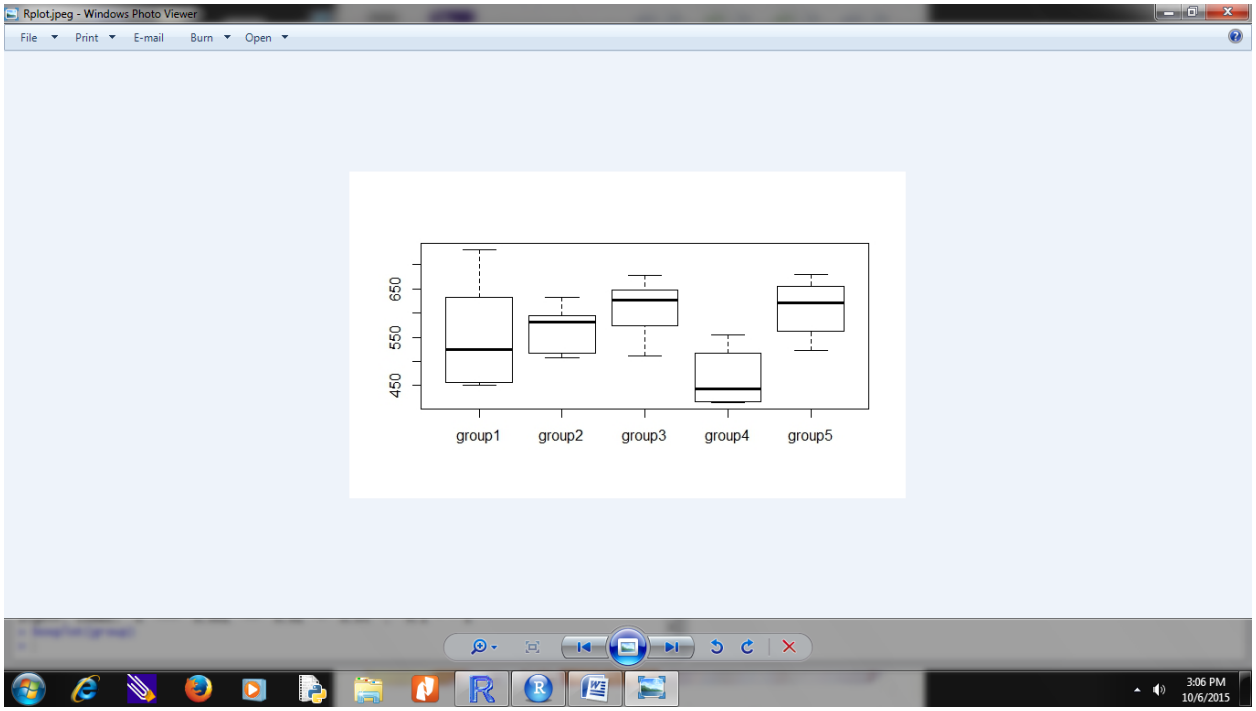
Syntax	Description
read.table(file,header=TRUE)	The name of the *file* which the data are to be read form. *header* a logical value indicating whether the file contains the names of the variables as its first line.
file.choose()	Choose a file interactively from the directory.
as.matrix()	Converts its first argument into a matrix, the dimensions of which will be inferred from the input.
gl(n,k,n*k,labels=seq_len(n))	Generate factors by specifying the pattern of their levels.
aov(arg1~arg2)	Anova command. The first argument is always the dependent variable and the second is independent variable.
summary( )	It is a generic function to produce result summaries of the results of anova
Boxplot( )	Produce box and whisker plots of the given values.

CODE:

```
group1<-c(551,457,450,731,499,632)

group2<-c(595,580,508,583,633,517)

group3<-c(639,615,511,573,648,677)

group4<-c(417,449,517,438,415,555)

group5<-c(563,631,522,613,656,679)

group<-data.frame(cbind(group1,group2,group3,group4,group5))

summary(group)

stgr<-stack(group);

crd<-aov(values~ind,data=stgr)

summary(crd)

boxplot(group)
```

OUTPUT

```
summary(group)
    group1          group2          group3          group4          group5
 Min.   :450.0   Min.   :508.0   Min.   :511.0   Min.   :415.0   Min.   :522.0
 1st Qu.:467.5   1st Qu.:532.8   1st Qu.:583.5   1st Qu.:422.2   1st Qu.:575.5
 Median :525.0   Median :581.5   Median :627.0   Median :443.5   Median :622.0
 Mean   :553.3   Mean   :569.3   Mean   :610.5   Mean   :465.2   Mean   :610.7
 3rd Qu.:611.8   3rd Qu.:592.0   3rd Qu.:645.8   3rd Qu.:500.0   3rd Qu.:649.8
 Max.   :731.0   Max.   :633.0   Max.   :677.0   Max.   :555.0   Max.   :679.0
summary(crd)
            Df Sum Sq Mean Sq F value  Pr(>F)
ind          4  85356   21339   4.302 0.00875 **
Residuals   25 124020    4961
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```



INFERENCE:

P value 0.00875 is less than the alpha value (0.05). Therefore reject the null hypothesis. There is significant difference between groups.

Problem

1. Suppose the following table represents the sales figures of the 3 new menu items in the 18 restaurants after a week of test marketing. At .05 level of significance, test whether the mean sales volume for the 3 new menu items are all equal.

Item1	Item2	Item3
22	52	16
42	33	24
44	8	19
52	47	18
45	43	34
37	32	39

10. RANDOMIZED BLOCK DESIGN

AIM:

To find the ANOVA using RBD to test the null hypotheses against alternative hypotheses with level of significance , alpha=0.05.

MODEL:

$Y_{ij} = \mu + \alpha_i + \beta_j + \in_{ij}$ where $\in_{ij}$ is the deviation of the jth observation of the ith sample from the corresponding treatment mean (ie) random error, Y_{ij} jth observation from the ith treatment, μ is the grand mean , α_i is the effect of ith treatment and β_j is the effect of the jth block .

PROCEDURE:

Null hypothesis I (for treatment means)

The null hypothesis I is that the k treatment means are equal against the alternative that at least one of α_i is not equal to zero.

Null hypothesis II (for block means)

The null hypothesis II is that the b block means are equal against the alternative that at least one of β_j is not equal to zero.

The null hypothesis I is rejected if $P = P\{F[k-1,(k-1)(b-1)] > F\}$ *or* $\Pr(>F)$ is less than alpha value or else null hypothesis I is accepted.

The null hypothesis II is rejected if $P = P\{F[b-1,(k-1)(b-1)] > F\}$ *or* $\Pr(>F)$ is less than alpha value or else null hypothesis II is accepted.

SYNTAX USED

Syntax	Description
read.table(file,header=TRUE)	The name of the *file* which the data are to be read form. *header* a logical value indicating whether the file contains the names of the variables as its first line.
file.choose()	Choose a file interactively from the directory.
as.matrix()	Converts its first argument into a matrix, the dimensions of which will be inferred from the input.
gl(n,k,n*k,labels=seq_len(n))	Generate factors by specifying the pattern of their levels.

aov(arg1~arg2)	Anova command. The first argument is always the dependent variable and the second is independent variable.
summary( )	It is a generic function to produce result summaries of the results of anova
par(mfrow=c(n,k))	To create a matrix of n rows and k columns plots
interaction.plot()	Plots the mean of the response for two way combinations of factors.
plot( )	Generic function for plotting objects.

PROBLEM:

Four different machines M1, M2, M3 and M4 are being considered for the assembling of a particular product. It was decided that six different operators would be used in a randomized block experiment to compare the machines. The machines were assigned in a random order to each operator. The operation of the machines requires physical dexterity , and it was anticipated that there would be a difference among the operators in the speed with which they operated the machines. The amounts of time (in seconds) required to assemble the product are shown in the table below. Test the hypothesis H0 at the 0.05 level of significance, that the machines perform at the same mean rate of speed and there is no significance difference between the performances of the operators. (See the attached data file named as "data2.txt")

	Operator					
Machine	1	2	3	4	5	6
1	42.5	39.3	39.6	39.9	42.9	43.6
2	39.8	40.1	40.5	42.3	42.5	43.1
3	40.2	40.5	41.3	43.4	44.9	45.1
4	41.3	42.2	43.5	44.2	45.9	42.3

Solution:

```
data<-read.table(file.choose(),header=TRUE)
time=c(t(as.matrix(data)))
f=c("Oper1","Oper2","Oper3","Oper4","Oper5","Oper6")
g=c("M1","M2","M3","M4")
k=ncol(data)
n=nrow(data)
Operators=gl(k,1,n*k,factor(f))
Machines=gl(n,k,n*k,factor(g))
anova=aov(time ~ Machines + Operators)
summary(anova)
```

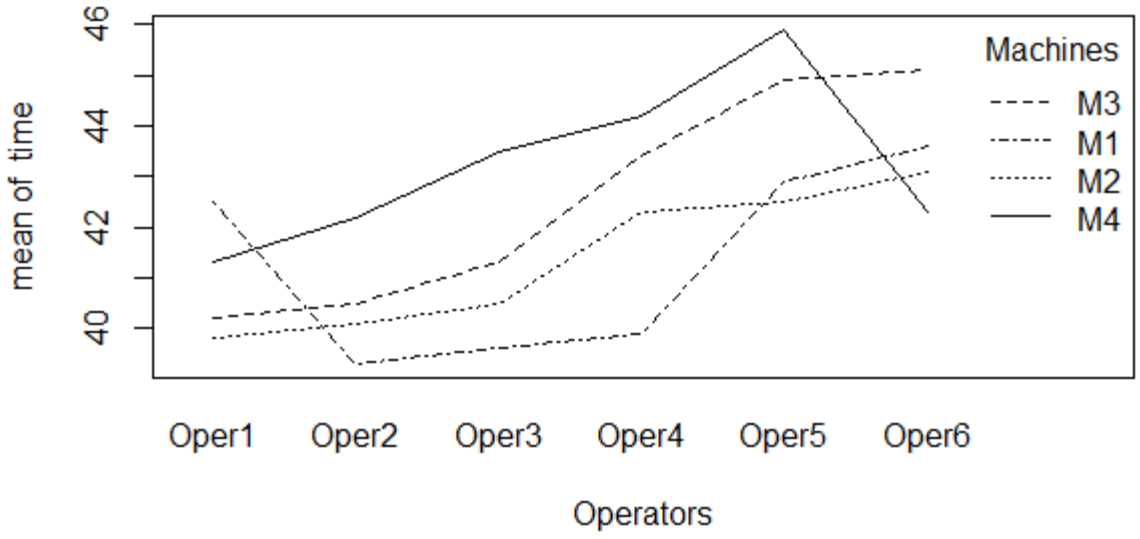
interaction.plot(Operators,Machines,time)
par(mfrow=c(1,2))
plot(time~Machines+Operators,main="Product time")

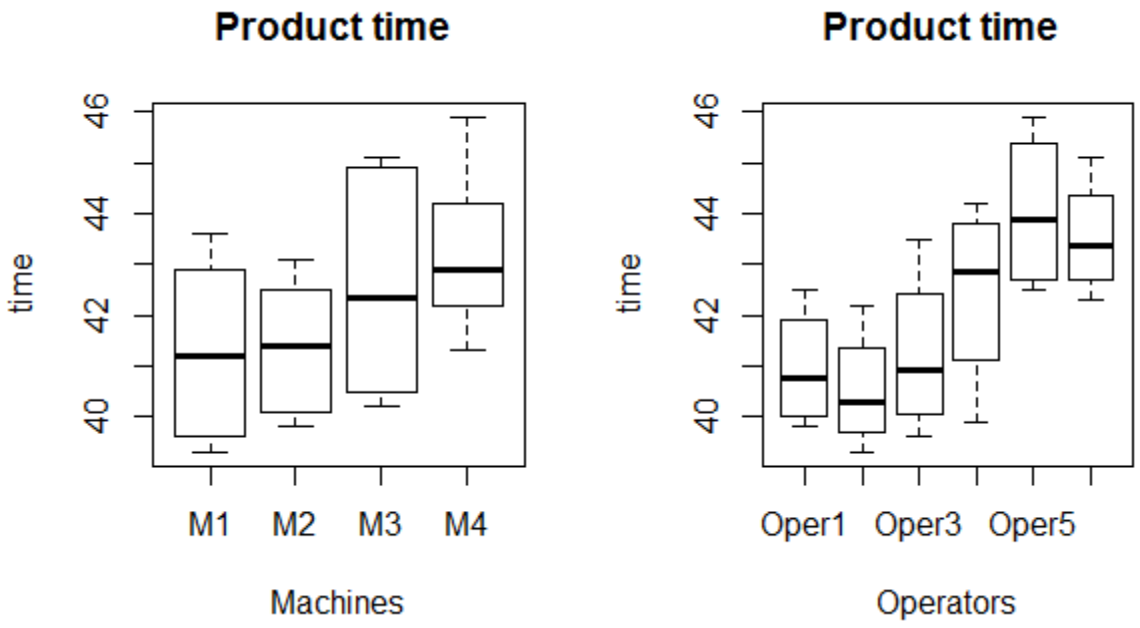
OUTPUT

summary(anova)

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
Machines	3	15.92	5.308	3.339	0.04790	*
Operators	5	42.09	8.417	5.294	0.00533	**
Residuals	15	23.85	1.590			

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





INFERENCE:

P value 0.04790 is less than the alpha value (0.05). Therefore reject the null hypothesis. There is significant difference between Machines. P-value 0.00533 is less than the alpha value 0.05, so reject the null hypothesis and conclude that there is some significant difference between operators also.

Problem

1. The following data represents the final grades obtained by five students in Mathematics, English, French and Biology. Test the hypothesis that the courses are of equal difficulty using the P-value in your conclusions and discuss your findings.

	Subject			
Student	Mathematics	English	French	Biology
1	68	57	73	61
2	83	94	91	86
3	72	81	63	59
4	55	73	77	66
5	92	68	75	87