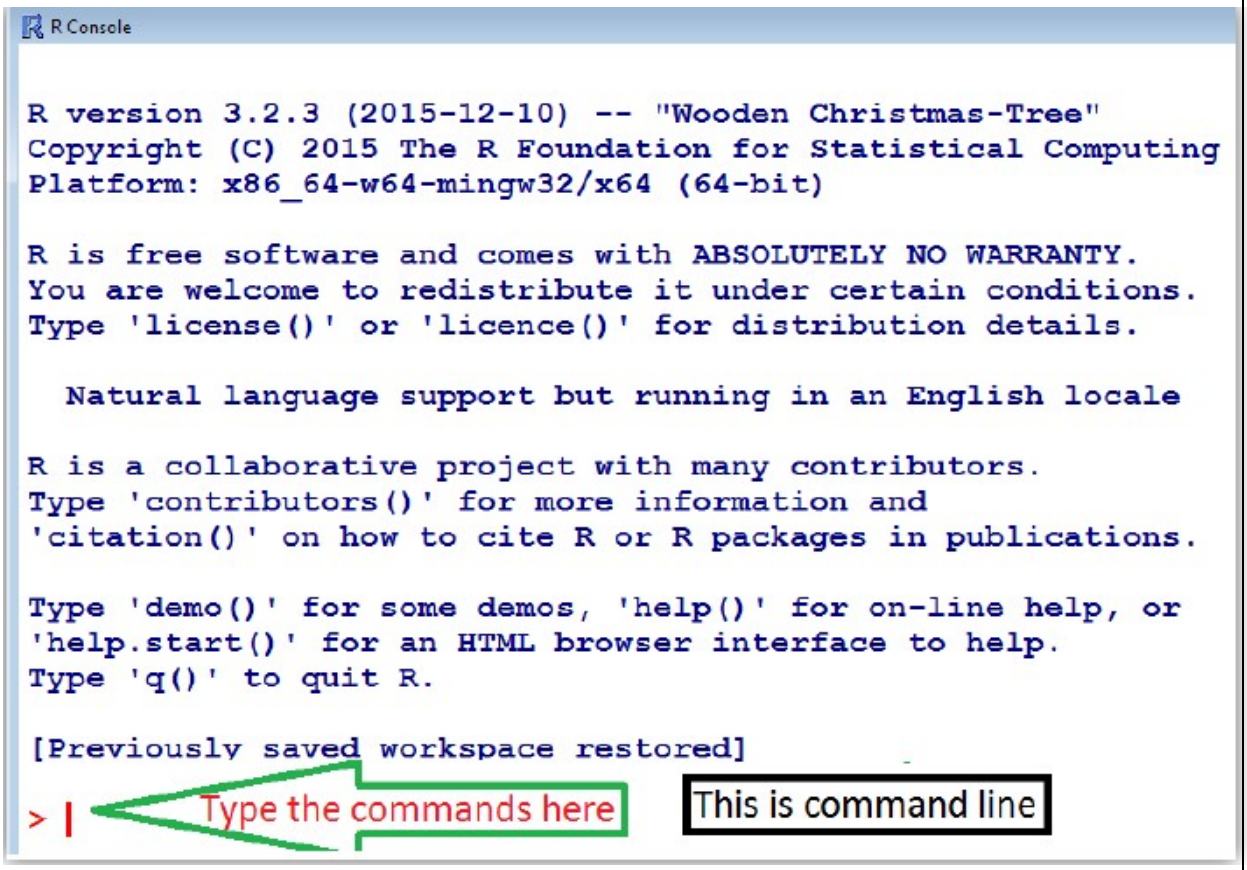
Introduction to R Language

- ✓ R is a computer language for carrying out statistical computations.
- ✓ R provides a wide variety of statistical (linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering, …) and graphical techniques, and is highly extensible.
- ✓ 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'.
- ✓ 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.
- ✓ The R Project is an international collaboration of researchers in statistical computing.
- ✓ It is a GNU project which is similar to the S language and environment which was developed at Bell Laboratories (formerly AT&T, now Lucent Technologies)by John Chambers and colleagues.
- ✓ R is available as Free Software under the terms of the Free Software Foundation's GNU General Public License in source code form.
- ✓ R, SAS, and SPSS are three statistical languages. Of these three statistical languages, R is the only an open source. SAS is the most important private software business in the world. SPSS is now overseen by IBM. R Programming is extensible and hence, R groups are noted for its energetic contributions.
- ✓ R offers plenty of options for loading external data, including Excel, Minitab, SAS and SPSS files.
- ✓ R can be downloaded from one of the mirror sites in http://cran.r-project.org/mirrors.html. You should pick your nearest location.s.

COMMAND LINE AND SCRIPT FILE

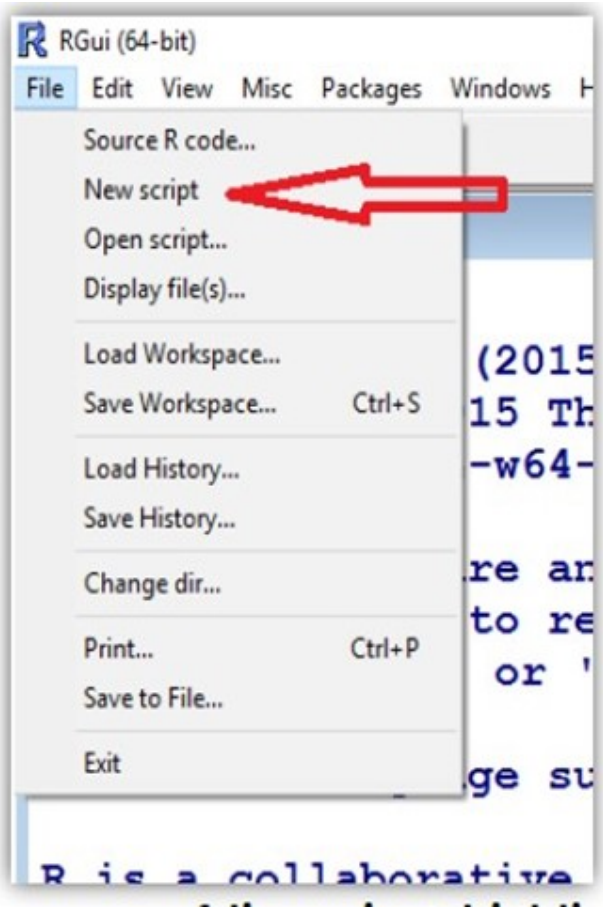


- Execution of commands in R is not menu driven. (Not like Clicking over buttons to get outcome)
- We need to type the commands
- Single line and multi line commands are possible to write.
- When writing multi-line programs it is useful to use a text editor rather than execute every line directly at the command line.

SCRIPT FILE :

At this point R will open a window entitled `Untitled-R Editor.`

We may type and edit in this.



If we want to execute a line or a group of lines, just highlight them and press `Ctrl+R`.

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 calculation.

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

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

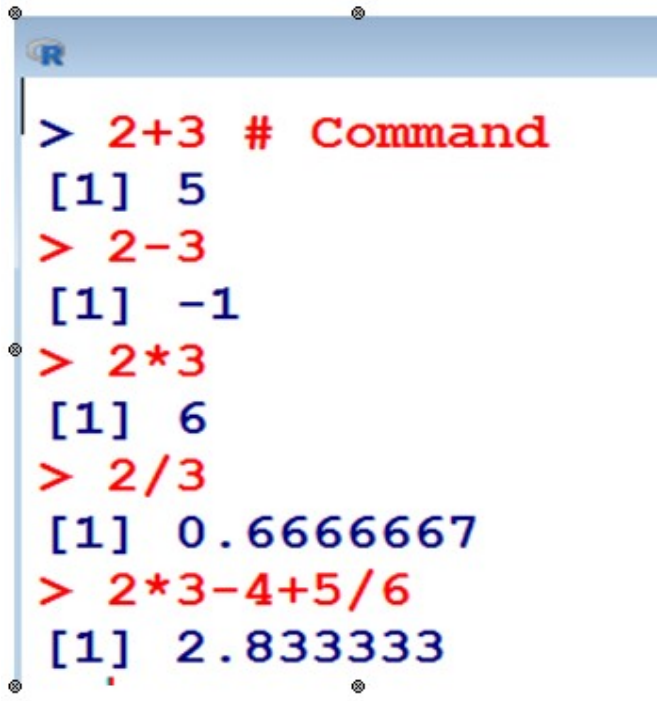
^	exponentiation
- +	unary minus and plus
:	sequence operator
%/%	integer division, remainder
* /	multiplication, division
+ -	addition, subtraction

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

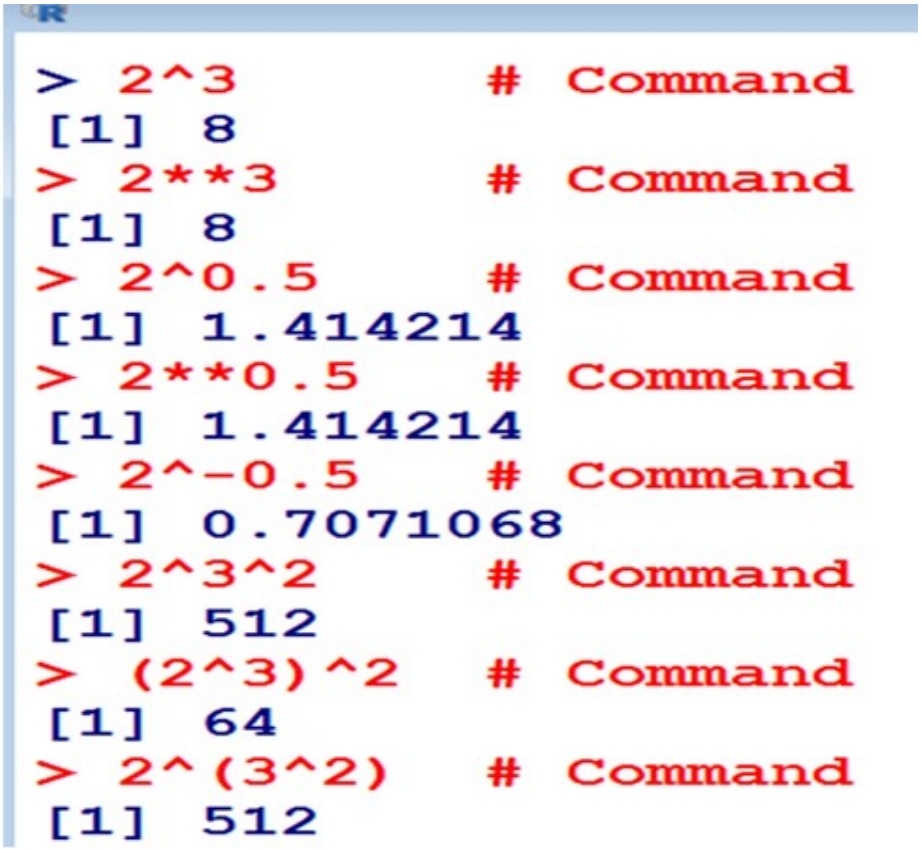
```
> # welcome to statistics class
> # 5+7
> 5+7
[1] 12
```

```
> 5# type 5 at the prompt
[1] 5
> 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 as a calculator

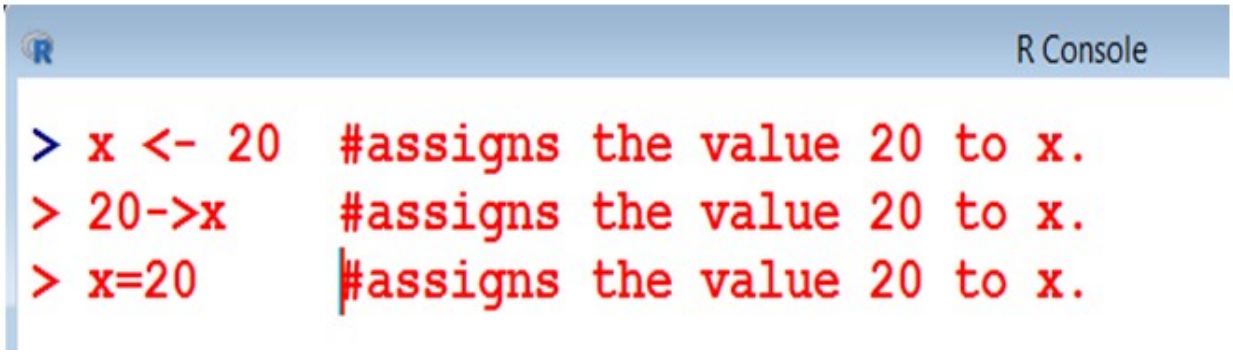


```
> 2+3 # Command
[1] 5
> 2-3
[1] -1
> 2*3
[1] 6
> 2/3
[1] 0.6666667
> 2*3-4+5/6
[1] 2.833333
```



```
> 2^3          # Command
[1] 8
> 2**3         # Command
[1] 8
> 2^0.5        # Command
[1] 1.414214
> 2**0.5       # Command
[1] 1.414214
> 2^-0.5       # Command
[1] 0.7071068
> 2^3^2        # Command
[1] 512
> (2^3)^2      # Command
[1] 64
> 2^(3^2)      # Command
[1] 512
```

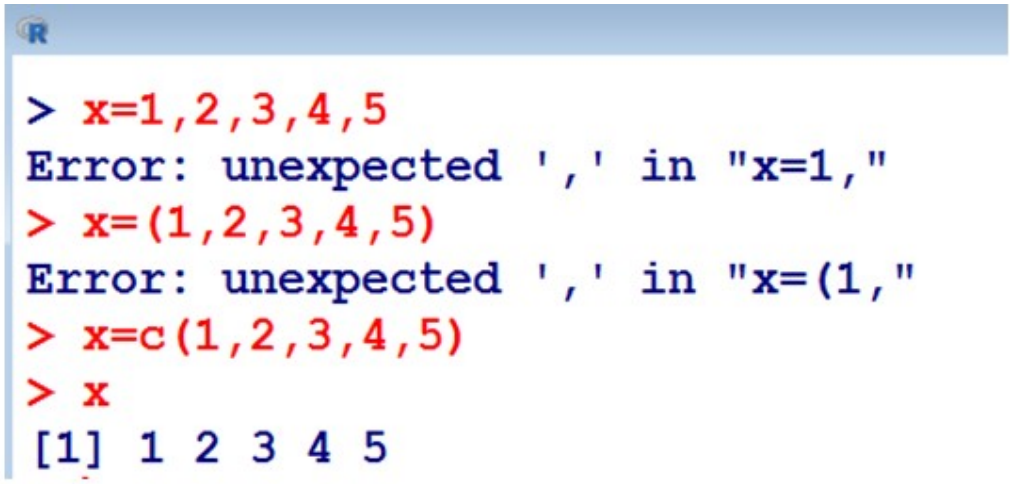
The assignment operators are : <- , -> ,=



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
```

Here the command c(1,2,3) combines the numbers 1,2 and 3 to a vector



- 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
```

```
R Console
> c(1,2,3,4,5)^2 #each vector is squared
[1]  1  4  9 16 25
> C(3,4,5,6)^c(2,3)
Error in C(3, 4, 5, 6) : object not interpretable as a factor
> # R is case sensitive C is Capital letter in above statment
> c(3,4,5,6)^c(2,3)
[1]   9  64  25 216
> # the above line is squared on rotation
> # the larger vector should be multiple of smaller vector
> c(3,4,5,6,2)^c(2,3)
[1]   9  64  25 216   4
Warning message:
In c(3, 4, 5, 6, 2)^c(2, 3) :
  longer object length is not a multiple of shorter object length
> c(3,4,5,6,2)^c(2,3,1)
[1]  9 64  5 36  8
Warning message:
In c(3, 4, 5, 6, 2)^c(2, 3, 1) :
  longer object length is not a multiple of shorter object length
> c(3,4,5,6,2,4)^c(2,3,1)
[1]  9 64  5 36  8  4
> c(2,3,4)+5
[1] 7 8 9
```

seq() function generates a sequence of numbers.

- ✓ Generate a sequence from -6 to 7:

```
> x <- seq(-6,7)
> x
[1] -6 -5 -4 -3 -2 -1 0 1 2 3 4 5 6 7
```

- From -6 till 7, step=2:

```
> x <- seq(-6,7,by=2)
> x
[1] -6 -4 -2 0 2 4 6
```

- Let's try smaller step:

```
> x <- seq(-2,2,by=0.3)
> x
[1] -2.0 -1.7 -1.4 -1.1 -0.8 -0.5 -0.2 0.1 0.4
 0.7 1.0 1.3 1.6 1.9
```

- Suppose we do not know the step, but we want 10 evenly distributed numbers from -2 to 2:

```
> seq(-2,2,length.out=10)
 [1] -2.0000000 -1.5555556 -1.1111111 -0.6666667 -0.2222222 0.2222222
[7] 0.6666667 1.1111111 1.5555556 2.0000000
```

- Generate a sequence from 1 to 10, quick version:

```
> x <- seq(10)
> x
 [1]  1  2  3  4  5  6  7  8  9 10
```

- **R rep Function**

rep() function replicates the values in x.

```
>x <- rep(1:5)
[1] 1 2 3 4 5
```

Repeat 1 -5 two times:

```
>x <- rep(1:5,2)
 [1] 1 2 3 4 5 1 2 3 4 5
```

Overview of further functions

`abs()`	Absolute value
`sqrt()`	Square root
`round(), floor(), ceiling()`	Rounding, up and down
`sum(), prod()`	Sum and product
`log(), log10(), log2()`	Logarithms
`exp()`	Exponential function
`sin(), cos(), tan(), asin(), acos(), atan()`	Trigonometric functions
`sinh(), cosh(), tanh(), asinh(), acosh(), atanh()`	Hyperbolic functions

```
> abs(-4)
[1] 4
> abs(c(-1,-2,-3,4,5))
[1] 1 2 3 4 5
> sqrt(4)
[1] 2
> sqrt(c(4,9,16,25))
[1] 2 3 4 5
> sum(c(2,3,5,7))
[1] 17
> prod(c(2,3,5,7))
[1] 210
> round(1.23)
[1] 1
> round(1.83)
[1] 2
```

- ***length()*** *function gets or sets the length of a vector (list) or other objects.*

Get vector length:

```
>x <- c(1,2,5,4,6,1,22,1)
>length(x)
[1] 8
```

Set vector length:

```
>length(x) <- 4
>x
[1] 1 2 5 4
```

R Matrix

R matrix is a two dimensional array. R has a lot of operator and functions that make matrix handling very convenient.

Matrix assignment:

```
>A <- matrix(c(3,5,7,1,9,4),nrow=3,ncol=2,byrow=TRUE)
>A
     [,1] [,2]
[1,]    3    5
[2,]    7    1
[3,]    9    4
```

Matrix row and column count:

```
>rA <- nrow(A)
>rA
[1] 3
>cA <- ncol(A)
>cA
[1] 2
```

t(A) function returns a transposed matrix of A:

```
>B <- t(A)
>B
     [,1] [,2] [,3]
[1,]    3    7    9
[2,]    5    1    4
```

Matrix multplication:

```
C <- A * A
C
```

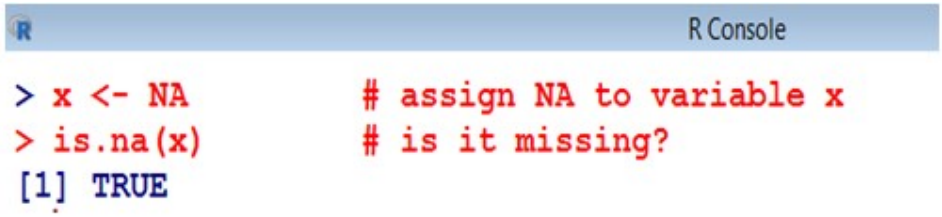
```
     [,1] [,2]
[1,]    9   25
[2,]   49    1
[3,]   81   16
```

Matrix Addition:

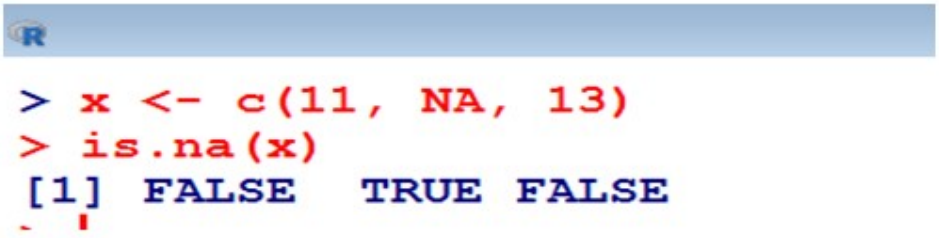
```
>C <- A + A
>C
     [,1] [,2]
[1,]    6   10
[2,]   14    2
[3,]   18    8
```

Missing Data

R represents missing observations through the data value NA We can detect missing values using is.na



Now try a vector to know if any value is missing?



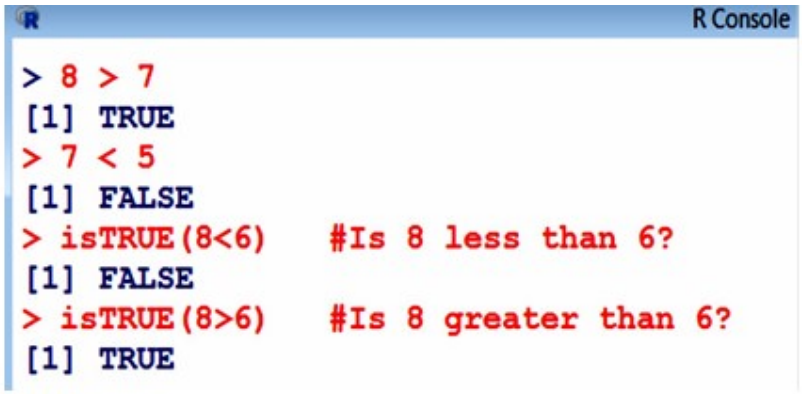
Logical Operators and Comparisons

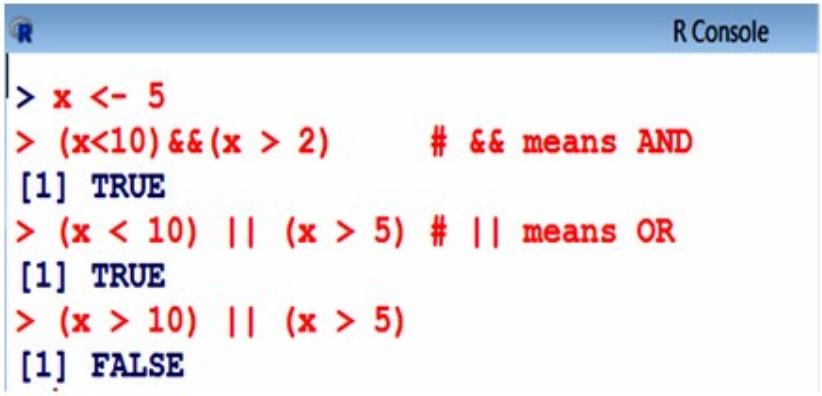
Operator	Executions
>	Greater than
>=	Greater than or equal
<	Less than
<=	Less than or equal
==	Exactly equal to
!=	Not equal to
!	Negation (not)

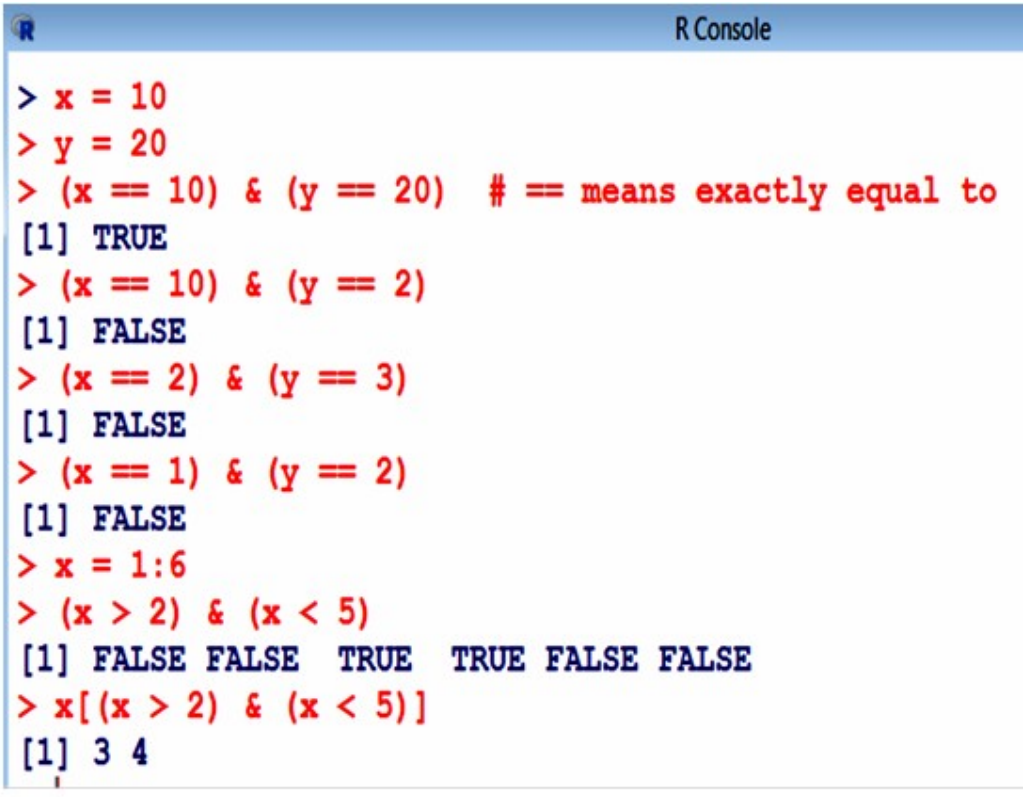
TRUE OR FALSE ARE RESERVED WORDS DENOTING LOGICAL CONSTANTS

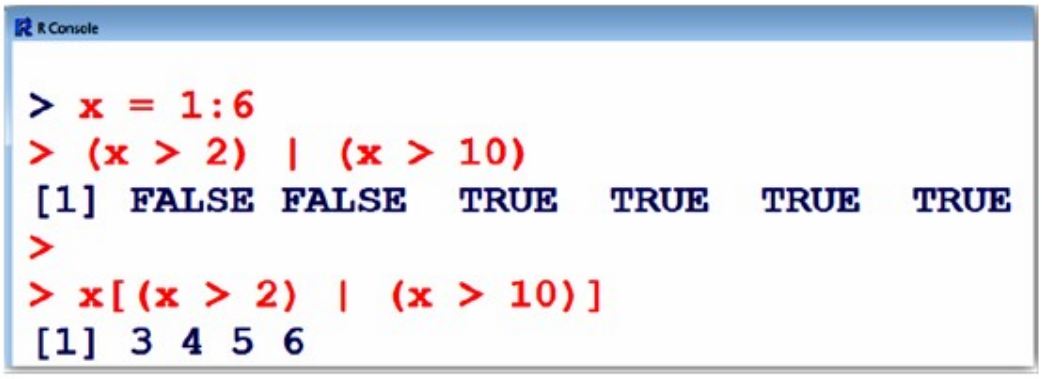
Operator	Executions
&, &&	and
|, ||	or

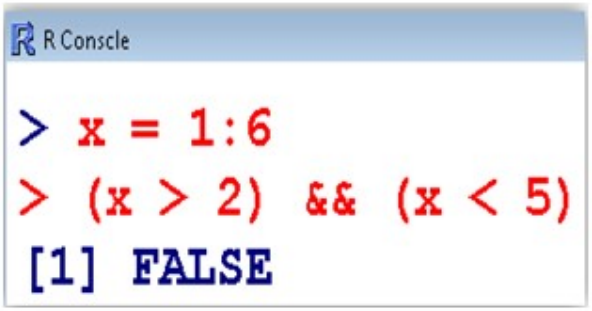
Operator	Executions
xor()	either... or (exclusive)
isTRUE(x)	test if x is TRUE
TRUE	true
FALSE	false


R Console
> 8 > 7
[1] TRUE
> 7 < 5
[1] FALSE
> isTRUE(8<6)   #Is 8 less than 6?
[1] FALSE
> isTRUE(8>6)   #Is 8 greater than 6?
[1] TRUE


R Console
> x <- 5
> (x<10)&&(x > 2)   # && means AND
[1] TRUE
> (x < 10) || (x > 5) # || means OR
[1] TRUE
> (x > 10) || (x > 5)
[1] FALSE


R Console
> x = 10
> y = 20
> (x == 10) & (y == 20) # == means exactly equal to
[1] TRUE
> (x == 10) & (y == 2)
[1] FALSE
> (x == 2) & (y == 3)
[1] FALSE
> (x == 1) & (y == 2)
[1] FALSE
> x = 1:6
> (x > 2) & (x < 5)
[1] FALSE FALSE TRUE TRUE FALSE FALSE
> x[(x > 2) & (x < 5)]
[1] 3 4


R Console
> x = 1:6
> (x > 2) | (x > 10)
[1] FALSE FALSE TRUE TRUE TRUE TRUE
>
> x[(x > 2) | (x > 10)]
[1] 3 4 5 6


R Conscle
> x = 1:6
> (x > 2) && (x < 5)
[1] FALSE


R Console
> (x[1] > 2) & (x[1] < 5)
[1] FALSE

Experiment Lab-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

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

>Gender

[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,Gender,status)

>empinfo

	empid	age	Gender	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$Gender=factor(empinfo$Gender,labels=c("male","female"))

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

>empinfo

```
   empid age Gender  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

```
> Genderm=subset(empinfo,empinfo$Gender=='male')
> Genderm
   empid age Gender  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

```
> Genderf=subset(empinfo,empinfo$Gender=='female')
> Genderf
   empid age Gender  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          Gender      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(Genderm)
     empid            age          Gender      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(Genderf)
     empid            age          Gender      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 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 Gender

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

  male female
     8      7
```

2. For status

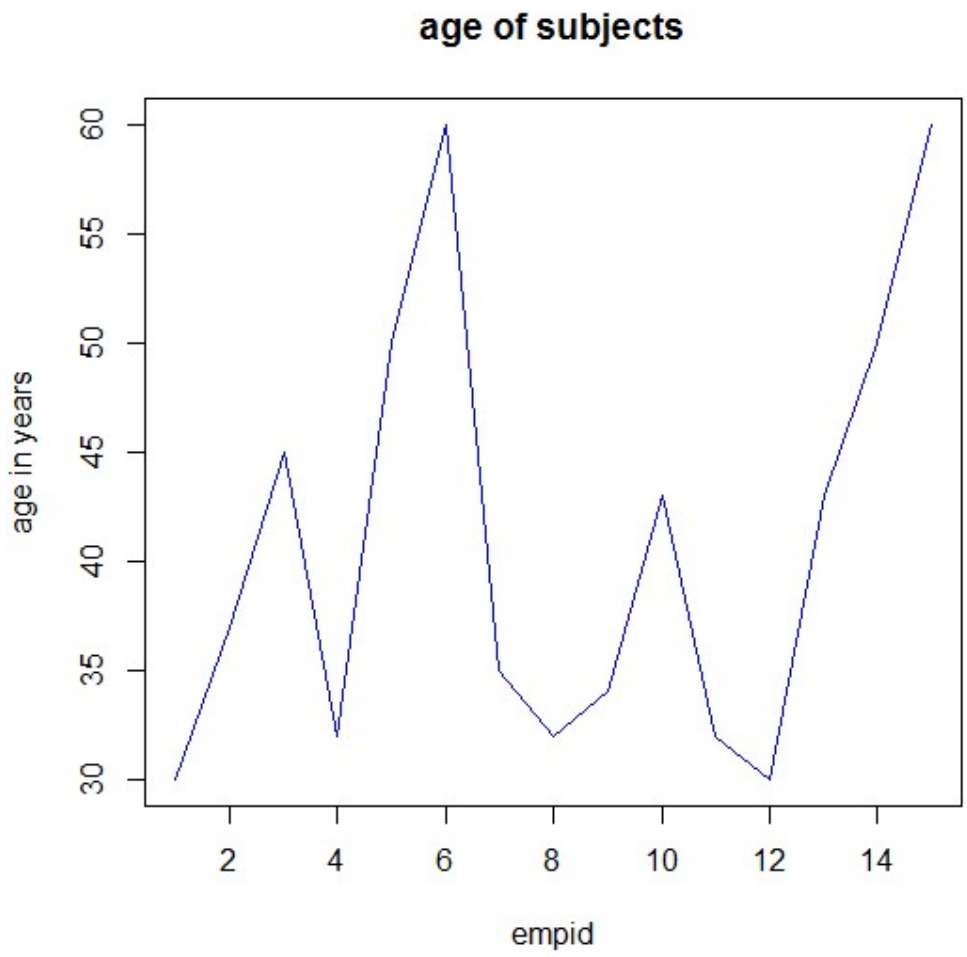
```
> table2=table(empinfo$status)
> table2

  staff faculty
      8       7
```

- Creating two-way table

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

        staff faculty
  male      4       4
  female    4       3
```

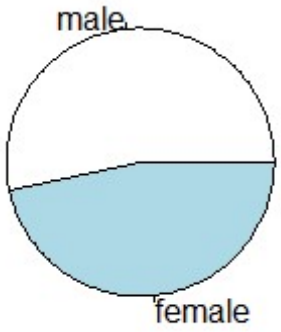
Gaphical reperesentation in R

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

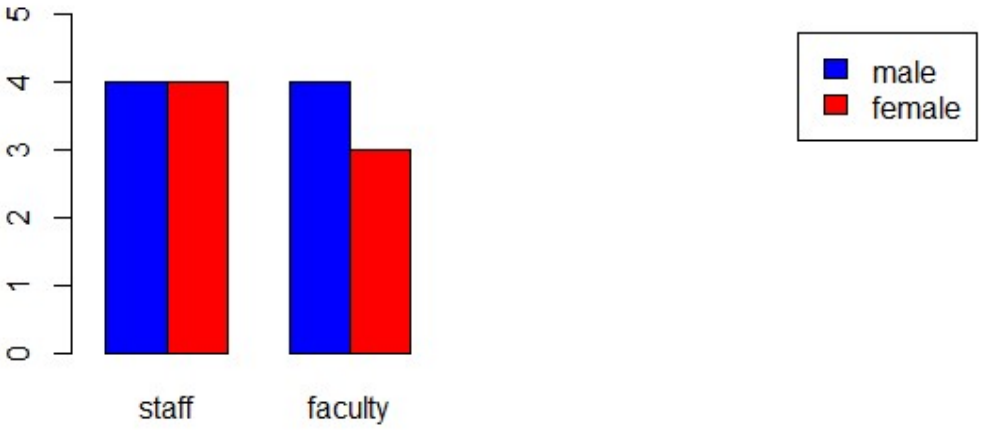


Pie Chart:-

```
> table4<-table(empinfo$Gender)
>pie(table4)
```

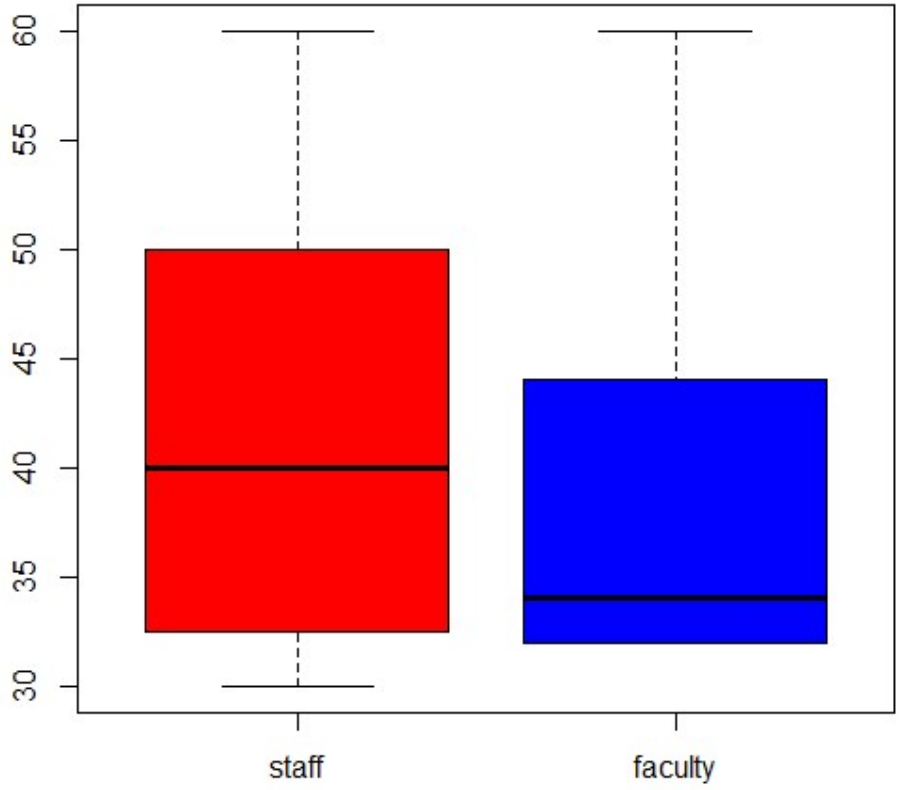



```
>table5=table(empinfo$Gender,empinfo$status)
>barplot(table5,beside=T,xlim=c(1,15),ylim=c(0,5),c
ol=c("blue","red"),legend=rownames(table5))
```



BOXPLOT:-

- boxplot(empinfo$age~empinfo$status,col=c("red","blue"))



Experiment1-Descriptive Statistics

Descriptive statistics is a set of math used to summarize data. Descriptive statistics can be distribution, central tendency, and dispersion of data.The distribution can be a normal distribution or binomial distribution. The central tendency can be mean, median, and mode. The dispersion or spreadness can be the range, interquartile range, variance, and standard deviation.In this session, you will import a CSV file, Excel file and you will perform basic data processing. I will explain descriptive statistics, central tendency measurements, dispersion measurements. You will look into how R programming can be used to calculate all these values.

What Is Descriptive Statistics?

Descriptive statistics summarizes the data and usually focuses on the distribution, the central tendency, and dispersion of the data. The distributions can be normal distribution, binomial distribution, and other distributions like Bernoulli distribution. Binomial distribution and normal distribution are the more popular and important distributions, especially normal distribution. When exploring data and many statistical tests, you will usually look for the normality of the data, which is how normal the data is or how likely it is that the data is normally distributed. The Central Limit Theorem states that the mean of a sample or subset of a distribution will be equal to the normal distribution mean when the sample size increases, regardless whether the sample is from a normal distribution. The central tendency, not the central limit theorem, is used to describe the data with respect to the center of the data. Central tendency can be the mean, median, and mode of the data. The dispersion describes the spread of the data, and dispersion can be the variance, standard deviation, and interquantile range. Descriptive statistics summarizes the data set, lets us have a feel and understanding of the data and variables, and allows us to decide or determine whether we should use inferential statistics to identify the relationship between data sets or use regression analysis to identify the relationships between variables.

Reading Data Files

R programming allow you to import a data set, which can be comma-separated values (CSV) file, Excel file, tab-separated file, JSON file, or others. Reading data into the R console or R is important, since you must have some data before you can do statistical computing and understand the data. Before you look into importing data into the R console, you must determine your workplace or work directory first. You should always set the current workspace directory to tell R the location of your current project folder. This allows for easier references to data files and scripts.

To print the current work directory, you use the getwd() function:

```
# get the current workspace location
print(getwd());
> print(getwd());
[1] "C:/Users/gohmi/Documents"

#set the current workspace location

setwd("D:/R"); #input your own file directory, for here
we use "D:/R"
> setwd("D:/R");
```

To get the new work directory location, you can use the getwd() function:

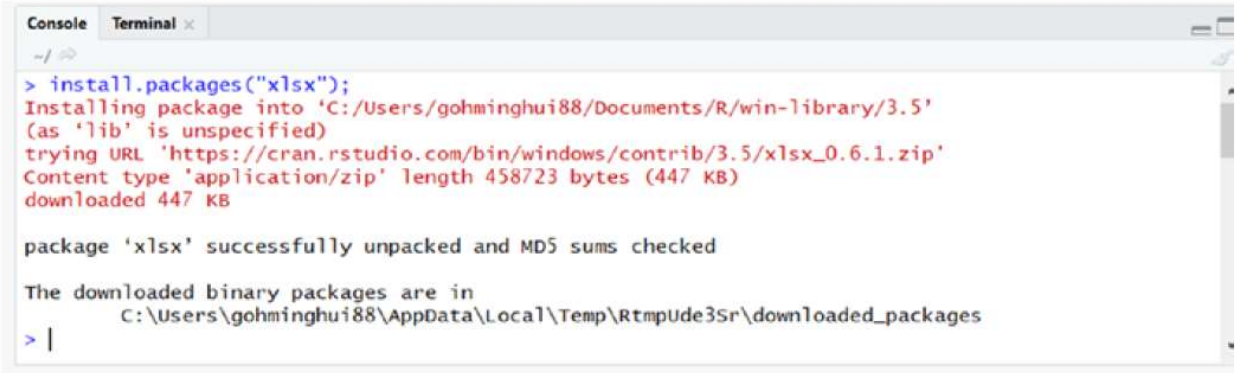
```
#get the new workspace
print(getwd());
> print(getwd());
[1] "D:/R"
```

You can put the data.csv data set into D:/R folder.

Reading an Excel File

The data set can also be in the Excel format or .xlsx format. To read an Excel file, you need to use the xlsx package. The xlsx package requires a Java runtime, so you must install it on your computer. To install the
xlsx package, go to the R console and type the following, also shown in Figure

```
> install.packages("xlsx");
```



To use the xlsx package, use the require() function:

```
> require("xlsx");
Loading required package: xlsx
```

To read the Excel file, you can use the read.xlsx() function:

> data <- read.xlsx(file="data.xlsx", 1);

file is the location of the Excel file. 1 refers to sheet number 1.To view the data variable, you can use the View() function or click the data variable in the Environment portion of RStudio, as shown in Figure.

To look for the documentation of read.xlsx(), you can use the following code.

> help(read.xlsx);

The data variable is of the data frame data type:
> class(data);
[1] "data.frame"

Writing an Excel File

To write a Excel file, you can use the write.xlsx() function:

> write.xlsx(data, file="data2.xlsx", sheetName="sheet1", col.names=TRUE, row.names=FALSE);
data is the variable of data frame type to export to Excel file, file is the file location or path, sheetName is the sheet name, and col.names and row.names are logical values to state whether to export with column names or row names. To view the documentation of the write.xlsx() function or any R function, you can use the help() function.

Basic Data Processing

After importing the data, you may need to do some simple data processing like selecting data, sorting data, filtering data, getting unique values, and removing missing values.

data=read.csv("C:/Users/dkalp/OneDrive/Desktop/spreadsheet.csv")

Mode, Median, Mean

Mean, median, and mode are the most common measures for central tendency. Central tendency is a measure that best summarizes the data and is a measure that is related to the center of the data set.

Mode

Mode is a value in data that has the highest frequency and is useful when the differences are non-numeric and seldom occur.

To get the mode in R, you start with data:

```
> A <- c(1, 2, 3, 4, 5, 5, 5, 6, 7, 8); #To get mode in a vector, you create a frequency table:
> y <- table(A);
> y;
A
1 2 3 4 5 6 7 8
1 1 1 1 3 1 1 1
```

You want to get the highest frequency, so you use the following to get the mode:

```
> names(y)[which(y==max(y))];
[1] "5"
```

Median

The median is the middle or midpoint of the data and is also the 50 percentile of the data. The median is affected by the outliers and skewness of the data. The median can be a better measurement for centrality than the mean if the data is skewed. The mean is the average, which is liable to be influenced by outliers, so median is a better measure when the data is skewed.

In R, to get the median, you use the median() function:

```
> A <- c(1, 2, 3, 4, 5, 5, 5, 6, 7, 8);
> median(A);
[1] 5
```

Mean

The mean is the average of the data. It is the sum of all data divided by the number of data points. The mean works best if the data is distributed in a normal distribution or distributed evenly. The mean represents the expected value if the distribution is random.

In R, to get the mean, you can use the mean() function:
> A <- c(1, 2, 3, 4, 5, 5, 5, 6, 7, 8);
> mean(A);
[1] 4.6

Handle NA Values with mean Function

A typical problem occurs when the data contains NAs. Let's modify our example vector to simulate such a situation:

>B=c(A,NA)
>B
[1] 1 2 3 4 5 5 5 6 7 8 NA

Our new example vector looks exactly the same as the first example vector, but this time with an NA value at the end. Let's see what happens when we apply the mean function as before:

>mean(B)
> [1] NA

The RStudio console returns NA – not as we wanted. Fortunately, the mean function comes with the na.rm (i.e. NA remove) option, which can be used to ignore NA values. Let's do this in practice:

>mean(B,na.rm=TRUE)
>[1] 4.6

As you can see, we get the same mean output as before.
Note:The na.rm option can also be used to ignore NaN or NULL values.

Problem1:Twenty students , graduates and undergraduates, were enrolled in a statistics course. Their ages were
18,19,19,19,19,20,20,20,20,20,21,21,21,21,22,23,24,27,30,36.
a) Find Mean and Median of all students
b) Find median age of all students under 25 years.
c) Find modal age of all student

R code:- >
x=c(18,19,19,19,19,20,20,20,20,20,21,21,21,21,22,23,24,27,30,36)
> mean(x) #mean
[1] 22
> median(x) #median
[1] 20.5
> y=x[x<25]
>median(y)
[1] 20
> xr=table(x) #mode
> mode=which(xr==max(xr))
> mode
20
3

Measures of central tendency for frequency table:-

Problem 2 : A survey of 25 faculty members is taken in a college to study their vocational mobility.They were asked the question “In addition to your present position ,at how many educational instistutes have served on the faculty?.Following is the frequency distribution of their responses .

X	*0*	*1*	*2*	*3*
f	*8*	*11*	*5*	*1*

Find mean and median of the distribution
R code:
> x=c(0,1,2,3)
> f=c(8,11,5,1)
> y=rep(x,f)
> mean=(sum(y))/(length(y)) #mean
> mean

[1] 0.96

> median(y) #median

[1] 1

Problem 3 : Compute mean ,median , 1st Quartile, 3rd Quartile and mode of for the following frequency Distribution:

Height in Cm	*145-150*	*150-155*	*155-160*	*160-165*	*165-170*	*170-175*	*175-180*	*180-185*
No. of Adult men	*4*	*6*	*28*	*58*	*64*	*30*	*5*	*5*

```
> x=seq(147.5,182.5,5)
> x
[1] 147.5 152.5 157.5 162.5 167.5 172.5 177.5 182.5
> f=c(4,6,28,58,64,30,5,5)
> mean=sum(x*f)/sum(f)
> mean
[1] 165.175

For Median:

> c=cumsum(f)
> cl=cumsum(f)
> cl
[1]   4  10  38  96 160 190 195 200
> N=sum(f)
> N
[1] 200
> ml=min(which(cl>N/2))
> ml
[1] 5
> h=5
> h
[1] 5
> fm=f[ml]
> fm
[1] 64
> cf=cl[ml-1]
> cf
[1] 96
> l=x[ml]-h/2
> l
[1] 165
> median=l+(((N/2)-cf)/fm)*h #median
> median
[1] 165.3125
```

To find Quartile 1:

```
> Q1=min(which(cl>N/4))
> Q1
[1] 4
> fq1=f[Q1]
> fq1
[1] 58
> cf1=cl[Q1-1]
> cf1
[1] 38
> l=x[Q1]-h/2
```

```
> l
[1] 160
> quartile1=l+(((N/4)-cf1)/fq1)*h
> quartile1
[1] 161.0345
```

To find Quartile 3:

```
> Q3=min(which(cl>3*N/4))
> Q3
[1] 5
> fq3=f[Q3]
> fq3
[1] 64
> cf2=cl[Q3-1]
> cf2
[1] 96
> l=x[Q3]-h/2
> l
[1] 165
> quartile3=l+(((3*N/4)-cf2)/fq3)*h
> quartile3
[1] 169.2188
```

Mode:

```
> m=which(f==max(f))
> m
[1] 5
> f0=f[m]
> f0
[1] 64
> f1=f[m-1]
> f1
[1] 58
> f2=f[m+1]
> f2
[1] 30
> l=x[m]-h/2
> l
[1] 165
> mode=l+((f0-f1)/(2*f0-f1-f2))*h
> mode
[1] 165.75
```

Range,Interquartile Range, Variance, Standard Deviation

Measures of variability are the measures of the spread of the data. Measures of variability can be range, interquartile range, variance, standard deviation, and more.

Range

The range is the difference between the largest and smallest points in the data.
To find the range in R, you use the *range()* function:

```
> A <- c(1, 2, 3, 4, 5, 5, 5, 6, 7, 8);
> range(A);
[1] 1 8
```

To get the difference between the max and the min, you can use

```
> A <- c(1, 2, 3, 4, 5, 5, 5, 6, 7, 8);
> res <- range(A);
> diff(res);
[1] 7
```

You can use the min() and max() functions to find the range also:

```
> A <- c(1, 2, 3, 4, 5, 5, 5, 6, 7, 8);
> min(A);
[1] 1
> max(A);
[1] 8
> max(A) - min(A);
[1] 7
```

To get the range for a data set:

```
> diff(res);
[1] 10.65222
```

Interquartile Range

The interquartile range is the measure of the difference between the 75 percentile or third quartile and the 25 percentile or first quartile.
To get the interquartile range, you can use the IQR() function:

```
> A <- c(1, 2, 3, 4, 5, 5, 5, 6, 7, 8);
> IQR(A);
[1] 2.5
```

You can get the quartiles by using the quantile() function:

```
> quantile(A);
0% 25% 50% 75% 100%
1.00 3.25 5.00 5.75 8.00
```

You can get the 25 and 75 percentiles:

```
> quantile(A, 0.25);
25%
3.25
```

```
> quantile(A, 0.75);
75%
5.75
```

The IQR() and quantile() functions can have NA values removed using na.rm = TRUE.

Range measures the maximum and minimum data value , and the interquartile range measures where the majority value is.

Example:
An entomologist studying morphological variation in species of mosquito recorded the following data on body length: 1.2,1.4,1.3,1.6,1.0,1.5,1.7,1.1,1.2,1.3.Compute all the measures of disersion.

```
> x=c(1.2,1.4,1.3,1.6,1.0,1.5,1.7,1.1,1.2,1.3)
> x
 [1] 1.2 1.4 1.3 1.6 1.0 1.5 1.7 1.1 1.2 1.3
> res=range(x)
> res
[1] 1.0 1.7
> diff(res)
[1] 0.7
> var(x)   # Variance
[1] 0.049
> sd(x)    # standard deviation
[1] 0.2213594
> quantile(x)
  0%   25%   50%   75%  100%
1.000 1.200 1.300 1.475 1.700

First Quartile  is 1.2
Second Quartile is 1.3
Third quartile is 1.475

> IQR(x)   # Inter quartile range
[1] 0.275
```

Mean deviation about Mean, Median and Mode:

```
> y=abs(x-mean(x))
> M1=sum(y)/length(y)  # mean deviation about Mean
> M1
[1] 0.176
> z=abs(x-median(x))
> M2=sum(z)/length(z)  # Mean deviation about median
> M2
[1] 0.17
Mean deviation about Mode # in this Problem ,it is a bi-model series (Mode is not
possible)
```

References

1. Biological data analysis, Tartu 2006/2007 (Tech.). (n.d.). Retrieved September 1, 2018, from www-1.ms.ut.ee/BDA/BDA4.pdf.

2. Calculate Standard Deviation. (n.d.). Retrieved from https://explorable.com/calculate-standard-deviation.

3. Descriptive Statistics. (n.d.). Retrieved from http://webspace.ship.edu/cgboer/descstats.html.

4.Descriptive statistics. (2018, August 22). Retrieved from https://en.wikipedia.org/wiki/Descriptive_statistics.

5.Donges, N. (2018, February 14). Intro to Descriptive Statistics – Towards Data Science. Retrieved from https://towardsdatascience.com/intro-to-descriptive-statistics-252e9c464ac9.

6. How to Make a Histogram with Basic R. (2017, May 04). Retrieved from www.r-bloggers.com/how-to-make-a-histogram-with-basic-r/.

Lab-5

(Challenging Experiment 3)

Find the linear correlation co-efficient and fit the regression line for future prediction. Applying simple linear regression model to real dataset; computing and interpreting the coefficient of determination	Conceptual understanding of Model fitting and investigate relationships between two variables within a regression framework	Learn to do future prediction with two variable

Correlation and Linear regression

Aim: Model fitting and investigating relationships between two variables within a regression framework.

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.

Problem:-

AGE GROUP	REPRESENTATIVE AGE	HOURS SPEND IN THE LOCAL LIBRARY
10-19	15	302.38
20-29	25	193.63
30-39	35	185.46
40-49	45	198.49

AGE GROUP	REPRESENTATIVE AGE	HOURS SPEND IN THE LOCAL LIBRARY
50-59	55	224.30
60-69	65	288.71

illustrate the relationship between the average age versus the time spent in the library, by using scatterplot.

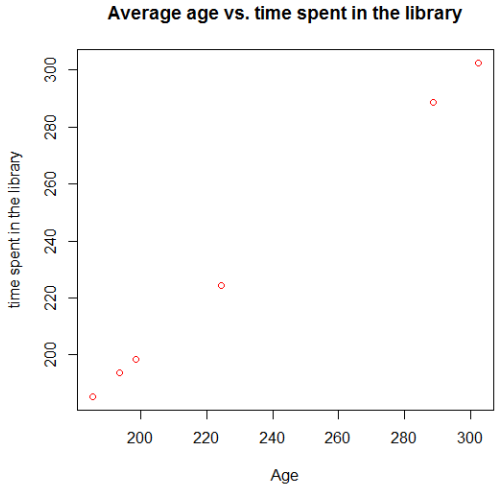
R code:-

>x <- c(302.38, 193.63, 185.46, 198.49, 224.30, 288.71)

> y <- c(302.38, 193.63, 185.46, 198.49, 224.30, 288.71)

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

OUTPUT:-



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 $\sigma_X \sigma_Y$

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

Consider a set of 'n' pairs of observations (X_1, Y_1), (X_2, Y_2), ... (X_n, Y_n) on two variables X and Y. Then we have, Covariance between X and Y

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)
>var(x)
[1] 15.33333
>var(y)
[1] 18.22222
>var(x,y)
[1] 13.66667
> r=var(x,y)/sqrt(var(x)*var(y))
> r
[1] 0.8176052
Or
>cor(x,y)
[1] 0.8176052
Or
>cor.test(x,y) Or
>cor.test(x,y,method="pearson")
Pearson's product-moment correlation
data:  x and y
t = 4.0164, df = 8, p-value = 0.003861
alternative hypothesis: true correlation is not equal to 0
```

95 percent confidence interval:

0.3874142 0.9554034

sample estimates:

cor

0.8176052

There is a Positive correlation between X and Y

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'

SPEARMAN'S RANK CORRELATION COFFICIENT FOR A DATA WITH AND WITHOUT TIED OBSERVATIONS:

Problem : 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 Scrore	*48*	*55*	*45*	*60*	*43*	*80*	*58*	*50*	*77*	*46*	*47*	*65*

Calculate rank correlation coefficient and comment on your result.

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")

Spearman's rank correlation rho

data: selection and proficielncy
S = 80, p-value = 0.01102

alternative hypothesis: true rho is not equal to 0
sample estimates:
rho
0.7202797

There is a positive correlation between selection and Proficiency

KENDALL'S COEFFICIENT OF CONCURRENT DEVIATIONS

The Kendall's coefficient of concurrent deviations is denoted by r_c and defined as

$$r_c = \pm\sqrt{\pm\left[\frac{2C - n}{n}\right]}$$

Where, C = Number of concurrent deviations or position signs of (D_X, D_Y);

n = Number of pairs of deviations

Problem: The following data gives the marks obtained by 12 students in statistics and computer science :

Students		*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*
Marks	*Statistics*	*55*	*40*	*70*	*60*	*62*	*73*	*65*	*65*	*20*	*35*	*46*	*50*
	Computer Science	*35*	*32*	*65*	*50*	*63*	*45*	*50*	*65*	*70*	*72*	*72*	*40*

Compute the coefficient of correlation by the method of concurrent deviations.
R code:

```
> statistics=c(55,40,70,60,62,73,65,65,20,35,46,50)
> mathematics=c(35,32,65,50,63,45,50,65,70,72,72,40)
>cor.test(statistics,mathematics,method="kendall")
Kendall's rank correlation tau
data:  statistics and mathematics
z = -0.27688, p-value = 0.7819
alternative hypothesis: true tau is not equal to 0
sample estimates:
      tau
```

-0.06250763

There is a negative correlation between mathematics and statistics

R^2 *(Coefficient of determination):-*

Code:

```
examdata=read.csv("C:\\Users\\aadmin\\Desktop\\mokesh\\examdata.csv")
examdata2 <- examData[, c("Exam", "Anxiety", "revise")]
cor(examdata2)
```

OUTPUT:-

	Exam	Anxiety	revise
Exam	1.0000000	-0.6381787	0.6281441
Anxiety	-0.6381787	1.0000000	-0.8190752
revise	0.6281441	-0.8190752	1.0000000

Interpretation:-

Provides a matrix of the correlation coefficients for the three variables. Each variable is perfectly correlated with itself (obviously) and so r = 1 along the diagonal of the table. Exam performance is negatively related to exam anxiety with a Pearson correlation coefficient of r = −.441. This is a reasonably big effect. Exam performance is positively related to the amount of time spent revising, with a coefficient of r = .397, which is also a reasonably big effect. Finally, exam anxiety appears to be negatively related to the time spent revising, r = −.709, which is a substantial effect size. In psychological terms, this all means that as anxiety about an exam increases, the percentage mark obtained in that exam decreases. Conversely, as the amount of time revising increases, the percentage obtained in the exam increases. Finally, as revision time increases, the student's anxiety about

the exam decreases. So there is a complex interrelationship between the three variables.

$R^{2:-}$

```
> examdata=read.csv("C:\\Users\\aadmin\\Desktop\\examdata.csv")
> examdata2 <- examdata[, c("Exam", "Anxiety", "revise")]
> cor(examdata2)^2     #coefficient of determination
             Exam   Anxiety    revise
Exam    1.0000000 0.4072769 0.3945650
Anxiety 0.4072769 1.0000000 0.6708793
revise  0.3945650 0.6708793 1.0000000
```

Interpretation:-

*Coefficient a step further by squaring it. The correlation coefficient squared (known as the **coefficient of determination**, R^2) is a measure of the amount of variability in one variable that is shared by the other. From the above we may look at the relationship between exam anxiety and exam performance. Exam performances vary from person to person because of any number of factors (different ability, different levels of preparation and so on). then we would have an estimate of how much variability exists in exam performances. We can then use* R^2 *to tell us how much of this variability is shared by exam anxiety. These two variables had a correlation of -0.6381787 and so the value of R^2 will be (-0.6381787) ^2 = 0.4072721.This value tells us how much of the variability in exam performance is shared by exam anxiety.*

If we convert this value into a percentage (multiply by 100) we can say that exam anxiety shares 40.7%1of the variability in exam performance. So, although exam anxiety was highly correlated with exam performance, it can account for only 40.7% of variation in exam scores. To put this value into perspective, this leaves 59.3 % of the variability still to be accounted for by other variables

Linear Regression Model

To draw conclusions about a population based on a regression analysis done on a sample, several assumptions must be true (see Berry, 1993):

***Variable types:** All predictor variables must be quantitative or categorical (with two categories), and the outcome variable must be quantitative, continuous and*

unbounded. By 'quantitative' I mean that they should be measured at the interval level and by 'unbounded' I mean that there should be no constraints on the variability of the outcome. If the outcome is a measure ranging from 1 to 10 yet the data collected vary between 3 and 7, then these data are constrained.

Non-zero variance: *The predictors should have some variation in value (i.e., they do not have variances of 0).*

No perfect multicollinearity: *There should be no perfect linear relationship between two or more of the predictors. So, the predictor variables should not correlate too highly (see section 7.7.2.4).*

Predictors are uncorrelated with 'external variables': *External variables are variables that haven't been included in the regression model which influence the outcome variable.9 These variables can be thought of as similar to the 'third variable' that was discussed with reference to correlation. This assumption means that there should be no external variables that correlate with any of the variables included in the regression model. Obviously, if external variables do correlate with the predictors, then the conclusions we draw from the model become unreliable (because other variables exist that can predict the outcome just as well).*

Homoscedasticity: *At each level of the predictor variable(s), the variance of the residual terms should be constant. This just means that the residuals at each level of the predictor(s) should have the same variance (homoscedasticity); when the variances are very unequal there is said to be heteroscedasticity (see section 5.7 as well).*

Independent errors: *For any two observations the residual terms should be uncorrelated (or independent). This eventuality is sometimes described as a lack of autocorrelation. This assumption can be tested with the Durbin–Watson test, which tests for serial correlations between errors. Specifically, it tests whether adjacent residuals are correlated. The test statistic can vary between 0 and 4, with a value of 2 meaning that the residuals are uncorrelated. A value greater than 2 indicates a negative correlation between adjacent residuals, whereas a value less than 2 indicates a positive correlation. The size of the Durbin–Watson statistic depends upon the number of predictors in the model and the number of observations. As a very conservative rule of thumb, values less than 1 or greater than 3 are definitely cause for concern; however, values closer to 2 may still be problematic depending on your sample and model. R also provides a p-value of the autocorrelation. Be very careful with the Durbin–Watson test, though, as it depends on the order of the data: if you reorder your data, you'll get a different value.*

Normally distributed errors: *It is assumed that the residuals in the model are random, normally distributed variables with a mean of 0. This assumption simply means that the differences between the model and the observed data are most frequently zero or very close to zero, and that differences much greater than zero happen only occasionally. Some people confuse this assumption with the idea that predictors have to be normally distributed. Predictors do not need to be normally distributed .*

Independence: *It is assumed that all of the values of the outcome variable are independent (in other words, each value of the outcome variable comes from a separate entity).*

Linearity: *The mean values of the outcome variable for each increment of the predictor(s) lie along a straight line. In plain English this means that it is assumed that the relationship we are modelling is a linear one. If we model a non-linear relationship using a linear model then this obviously limits the generalizability of the findings.*

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.

Answer:-

```
> weight=c(15,26,27,25,25.5,27,32,18,22,20,26,24)
> bmi=c(13.35,16.12,16.74,16.00,13.59,15.73,15.65,13.85,16.07,12.8,13.65,14.42)
> cor(weight,bmi)
[1] 0.5790235
> model<-lm(bmi~weight)
> summary.lm(model)

Call:
lm(formula = bmi ~ weight)

Residuals:
     Min       1Q   Median       3Q      Max
-1.52988 -0.75527  0.04426  0.95286  1.57397

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 10.73487    1.85405   5.790 0.000175 ***
weight       0.17096    0.07612   2.246 0.048524 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.155 on 10 degrees of freedom
Multiple R-squared:  0.3353,    Adjusted R-squared:  0.2688
F-statistic: 5.044 on 1 and 10 DF,  p-value: 0.04852
```

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.

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***

CODE:-

```
> height = c(175, 168, 170, 171, 169, 165, 165, 160, 180, 186)
> weight = c(80, 68, 72, 75, 70, 65, 62, 60, 85, 90)
> cor(height,weight)
[1] 0.9849472
> model = lm(formula = height ~ weight)
> model

Call:
lm(formula = height ~ weight)

Coefficients:
(Intercept)       weight
   115.2002       0.7662

> summary(model)

Call:
lm(formula = height ~ weight)

Residuals:
    Min      1Q  Median      3Q     Max
-1.6622 -0.9683 -0.1622  0.5679  2.2979

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 115.20021    3.48450   33.06 7.64e-10 ***
weight        0.76616    0.04754   16.12 2.21e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.405 on 8 degrees of freedom
Multiple R-squared:  0.9701,    Adjusted R-squared:  0.9664
F-statistic: 259.7 on 1 and 8 DF,  p-value: 2.206e-07
```

Interpretation:-

The regression equation is

Height=115.2002+0.7662 weight. Since the p- value of the test is 16.12 is greater than 0.05 we reject the hypothesis. Therefore the model we found is significant. The Multiple R-squared is the coefficient of determination. It provides a measure of how well future outcomes are likely to be predicted by the model. In this case the R suare value is 0.9701. Therefore 97.01% of data is well predicted.

Practice Problem:

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***

Let us carry out linear regression analysis for this data.

2. ***The success of a shopping center can be represented as a function of the distance (inmiles) from the center of the population and the number of clients (in hundreds of people) who will visit. The data is given in the table below:***

No. Customer(x)	***8***	***7***	***6***	***4***	***2***	***1***
Distance(y)	***15***	***19***	***25***	***23***	***34***	***40***

   a) ***Calculate the linear correlation coefficient***
   b) ***If the mall is located 2 miles from the center of the population, how many customers should the shopping center except?***
   c) ***To receive 500 customers, at what distance from the center of the population should the shopping centre be located?***

3. ***Find the correlation between Experience and Income. Also fit a regression equation and interpret the result.***

Exp 3a-Binomial and Poisson Distributions

The Binomial distribution

Consider the following circumstances (binomial scenario):

1. There are n trials.

2. The trials are independent.

3. On each trial, only two things can happen. We refer to these two events as success and failure.

4. The probability of success is the same on each trial. This probability is usually called p.

5. We count the total number of successes. This is a discrete random variable, which we denote by X, and which can take any value between 0 and n (inclusive).

• The random variable X is said to have a binomial distribution with parameters n and p; abbreviated

$$X \sim \text{Bin}(n, p)$$

- It is easy to show that if $X \sim \text{Bin}(n, p)$ then

$$\text{P}[X = k] = \binom{n}{k} p^k (1-p)^{n-k}$$

for $k = 0, 1, \ldots, n$.

- $\binom{n}{k}$ is the *binomial coefficient* and is the number of sequences of length n containing k successes.

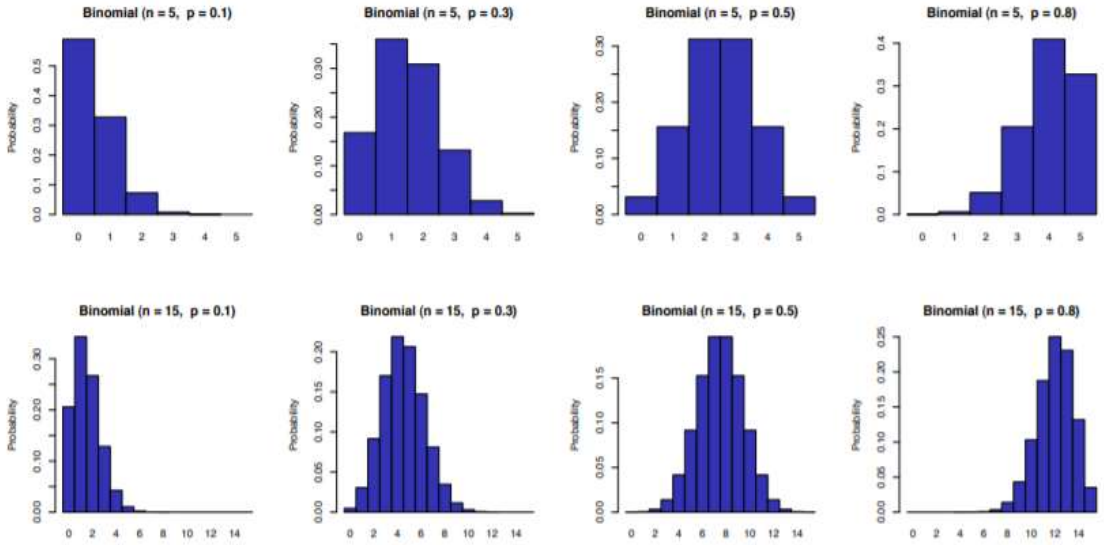
$$\binom{n}{k} = \frac{n!}{(n-k)!k!}$$

- The expectation and variance of X are given by

$$\text{E}[X] = np$$
$$\text{Var}[X] = np(1-p)$$

The Binomial Distribution: Example

The shape of the distribution depends on n and p.



R has four in-built functions to generate binomial distribution. They are described below.

```
dbinom(x, size, prob)
pbinom(x, size, prob)
qbinom(p, size, prob)
rbinom(n, size, prob)
```

Following is the description of the parameters used −

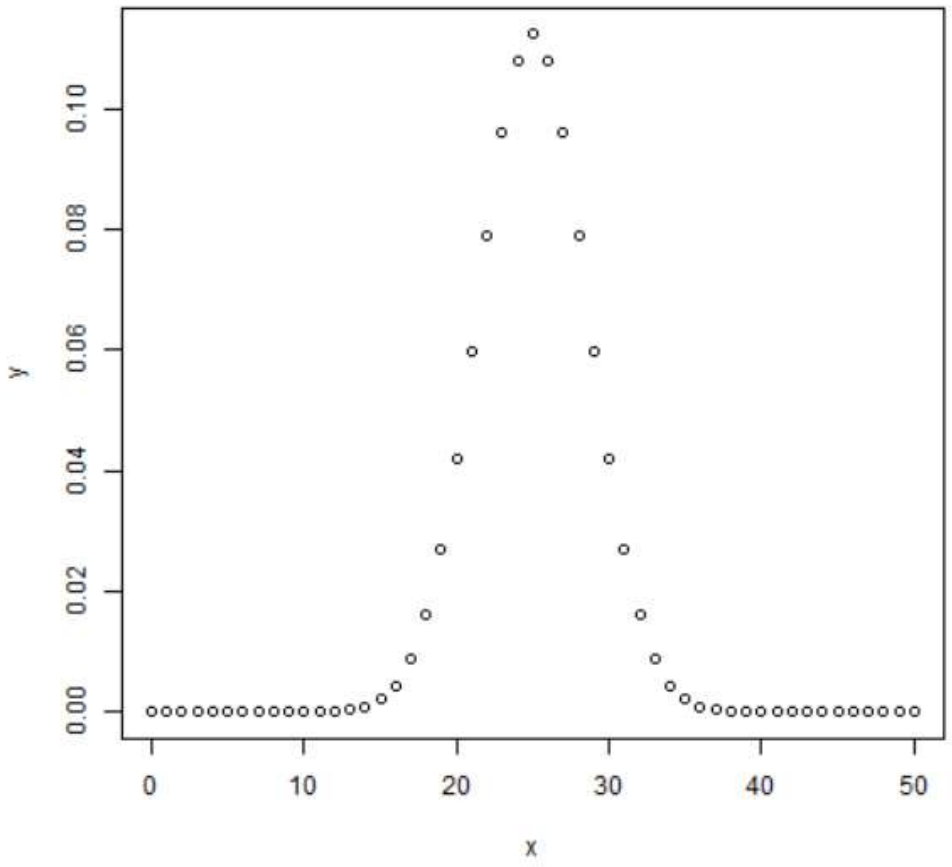
- **x** is a vector of numbers.
- **p** is a vector of probabilities.
- **n** is number of observations.
- **size** is the number of trials.
- **prob** is the probability of success of each trial.

dbinom()

This function gives the probability density distribution at each point.

```
# Create a sample of 50 numbers which are incremented by 1.
x <- seq(0,50,by = 1)

# Create the binomial distribution.
y <- dbinom(x,50,0.5)
```

```
# Give the chart file a name.
png(file = "dbinom.png")

# Plot the graph for this sample.
plot(x,y)

# Save the file.
dev.off()
```



pbinom()

This function gives the cumulative probability of an event. It is a single value representing the probability.

```
# Probability of getting 26 or less heads from 51 tosses of a coin.
x <- pbinom(26,51,0.5)

print(x)
```

[1] 0.610116

qbinom()

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

How many heads will have a probability of 0.25 will come out when a coin is tossed 51 times.

```
x <- qbinom(0.25,51,1/2)
print(x)
```
[1] 23

rbinom()

This function generates required number of random values of given probability from a given sample.

Find 8 random values from a sample of 150 with probability of 0.4.

```
x <- rbinom(8,150,.4)
print(x)
```
[1] 58 61 59 66 55 60 61 67

Example:

1. Let X~Bin(5,0.9). Find (a) $P(X \leq 4)$ and $P(X = 4)$

(a)> sum(dbinom(0:4,5,0.9))
[1] 0.40951

(b) > dbinom(4,5,0.9)
[1] 0.32805

2. The proportion of students wearing spectacles is 40%. Let X be the number of students wearing spectacles in a random sample of 10 students. Find

(a) $P(X \leq 2)$; (b) $P(2 \leq X < 5)$; (c) $P(X > 2)$;

(a) > sum(dbinom(0:2,10,0.4))
[1] 0.1672898
Or
>pbinom(2,10,0.4)

[1] 0.1672898

(b)> sum(dbinom(2:4,10,0.4))
[1] 0.5867459

(c) $P(X > 2) = 1 - P(X \leq 2)$
> 1-pbin
om(2,10,0.4)
[1] 0.8327102

3. If a committee has 7 members, find the probability of having more female me mbers than male members given that the probability of having a male or a femal e member is equal.

Sol: The probability of having a female member = 0.5
The probability of having a male member = 0.5
To have more female members, the number of females should be greater than or equal to 4.

```
> 1-pbinom(3,7,0.5)
[1] 0.5
```

4. In a box of switches it is known 10% of the switches are faulty. A technician is wiring 30 circuits, each of which needs one switch. What is the probability that (a) all 30 work, (b) at most 2 of the circuits do not work?

(a) Probability that all 30 work is $P(X = 30) = {}^{30}C_{30}(0.9)^{30}(0.1)^0 = 0.04239$

(b) The statement that "at most 2 circuits do not work" implies that 28, 29 or 30 work. That is $X \geq 28$

$$
\begin{aligned}
P(X \geq 28) &= P(X = 28) + P(X = 29) + P(X = 30) \\
P(X = 30) &= {}^{30}C_{30}(0.9)^{30}(0.1)^0 = 0.04239 \\
P(X = 29) &= {}^{30}C_{29}(0.9)^{29}(0.1)^1 = 0.14130 \\
P(X = 28) &= {}^{30}C_{28}(0.9)^{28}(0.1)^2 = 0.22766
\end{aligned}
$$

Hence $P(X \geq 28) = 0.41135$

```
> dbinom(30,30,0.9)
[1] 0.04239116
```

```
> 1-pbinom(27,30,0.9)
[1] 0.4113512
```

5. 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)

(i) # Exactly 2 defective

```
dbinom(2,20,0.10)
[1] 0.2851798
```

(ii) At least 2 defectives

```
1-pbinom(2,20,0.10)
[1] 0.3230732
```

(iii) Between 1 and 3 defectives (inclusive)

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

Poisson Distribution in R

We call it the distribution of rare events., a Poisson process is where DISCRETE events occur in a continuous, but finite interval of time or space in R

The following conditions must apply:

- For a small interval, the probability of the event occurring is proportional to the size of the interval.
- The probability of more than one occurrence in the small interval is negligible.
- Each occurrence must be independent of others and must be at random.
- The events are often defects, accidents or unusual natural happenings, such as an earthquake.
- The parameter for the Poisson distribution is a lambda. It is average or mean of occurrences over a given interval.
- The probability function is: for $x = 0,1.2,3$ ....

Difference between Binomial and Poisson Distribution in R

Binomial Distribution:

- Fixed no. of Trials (n) [10 pie throws], although, only two possible outcomes are possible.
- A probability of success is constant(p).
- Each trial is independent.
- Also, it predicts no.s of successes within a set no. of trials.
- We use it to test for independence.

Poisson Distribution

- Infinite no. of trials.
- Also, it has unlimited no. of outcomes possible.
- The mean of the distribution is the same for all intervals.
- No. of occurrence in any given interval independent of others.
- Also, it predicts no. of occurrences per unit, time, space.
- We use it to test for independence.

R-Code

- dpois(x, lambda) # **the probability of x successes in a period when the expected number of events is lambda**
- ppois(q, lambda) # **the cumulative probability of less than or equal to q successes**
- qpois(p, lambda) # **returns the value (quantile) at the specified cumulative probability (percentile) p**
- rpois(n, lambda) # **returns n random numbers from the Poisson distribution**

Practice problems:

1. What is P(X = 4) with lambda 2.6?

```
> dpois(4, lambda = 2.6)
[1] 0.1414218
```

2. What is $P(X \geq 2)$ with lambda 3?

> **1-ppois(2,3)**

[1] 0.5768099

2. Consider a computer system with Poisson job-arrival stream at an ` average of 2 per minute. Determine the probability that in any one-minute interval there will be

(i) 0 jobs
(ii) Exactly 3 jobs
(iii) at most 3 arrivals

Solution:

Job arrivals lambda =2

(i) No job arrivals

```
> dpois(0,2)
[1] 0.1353353
```

(ii) Exactly 3 jobs

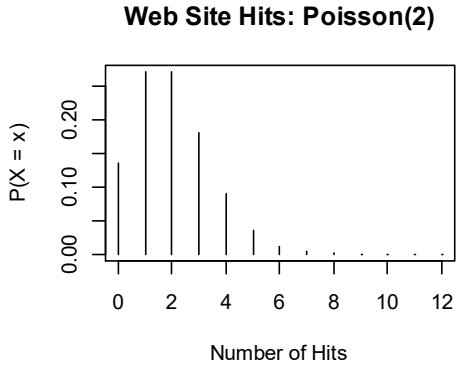
```
> dpois(3,2)
[1] 0.180447
```

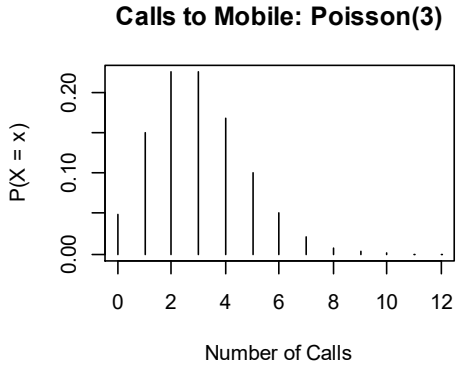
(iii) Atmost 3 job arrivals

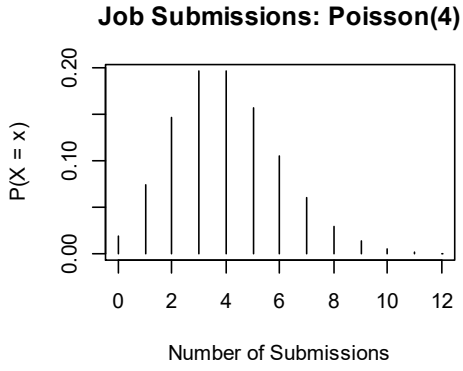
```
> ppois(3,2)
[1] 0.8571235
```

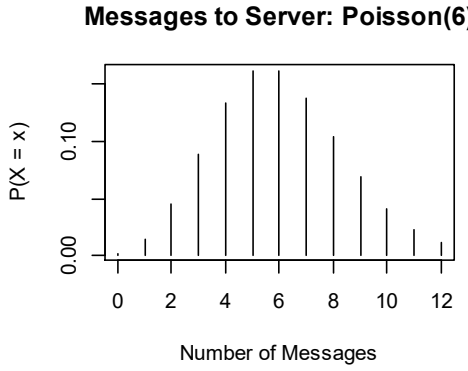
Poisson Probability Density Functions

```
par(mfrow = c(2,2))
# multiframe
x<-0:12 #look at the first 12 probabilities
plot (x, dpois(x, 2), xlab = "Number of Hits", ylab = "P(X = x)", type = "h
", main= "Web Site Hits: Poisson(2)")
plot (x, dpois(x, 3), xlab = "Number of Calls", ylab = "P(X = x)", type = "
h", main= "Calls to Mobile: Poisson(3)")
plot (x, dpois(x, 4), xlab = "Number of Submissions", ylab = "P(X = x)", t
ype = "h", main= "Job Submissions: Poisson(4)")
plot (x, dpois(x, 6), xlab = "Number of Messages", ylab = "P(X = x)", typ
e = "h", main= "Messages to Server: Poisson(6)")
```



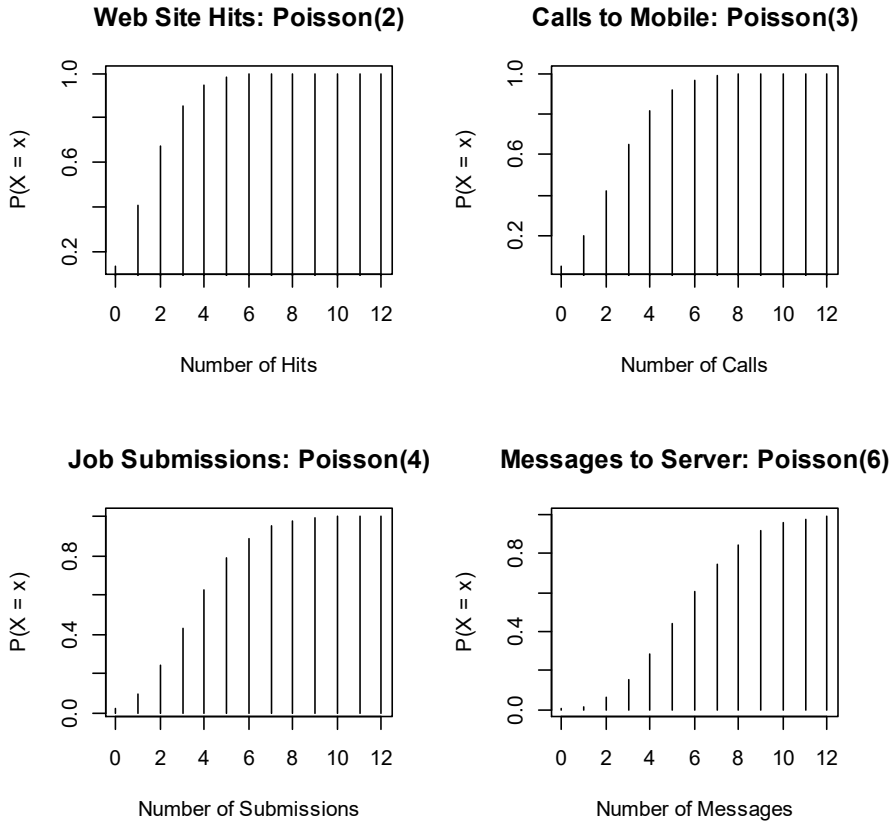






Poisson Cumulative Distribution Functions

```
par(mfrow = c(2,2))

# multiframe

x<-0:12   #look at the first 12 probabilities

plot (x, ppois(x, 2), xlab = "Number of Hits", ylab = "P(X = x)", type =
"h", main= "Web Site Hits: Poisson(2)")

plot (x, ppois(x, 3), xlab = "Number of Calls", ylab = "P(X = x)", type =
"h", main= "Calls to Mobile: Poisson(3)")

plot (x, ppois(x, 4), xlab = "Number of Submissions", ylab = "P(X = x)",
type = "h", main= "Job Submissions: Poisson(4)")

plot (x, ppois(x, 6), xlab = "Number of Messages", ylab = "P(X = x)",
type = "h", main= "Messages to Server: Poisson(6)")
```



Practice problems:

1. A recent national study showed that approximately 55.8% of college students have used Google as a source in at least one of their term papers. Let X equal the number of students in have used Google as a source:

a) Find the probability that X is equal to 17
b) Find the probability that X is at most 13.
c) Find the probability that X is bigger than 11.
d) Find the probability that X is at least 15.
e) Find the probability that X is between 16 and 19,
f) Give the mean of X
g) Give the variance of X.
h) Find E(4X + 51.324)

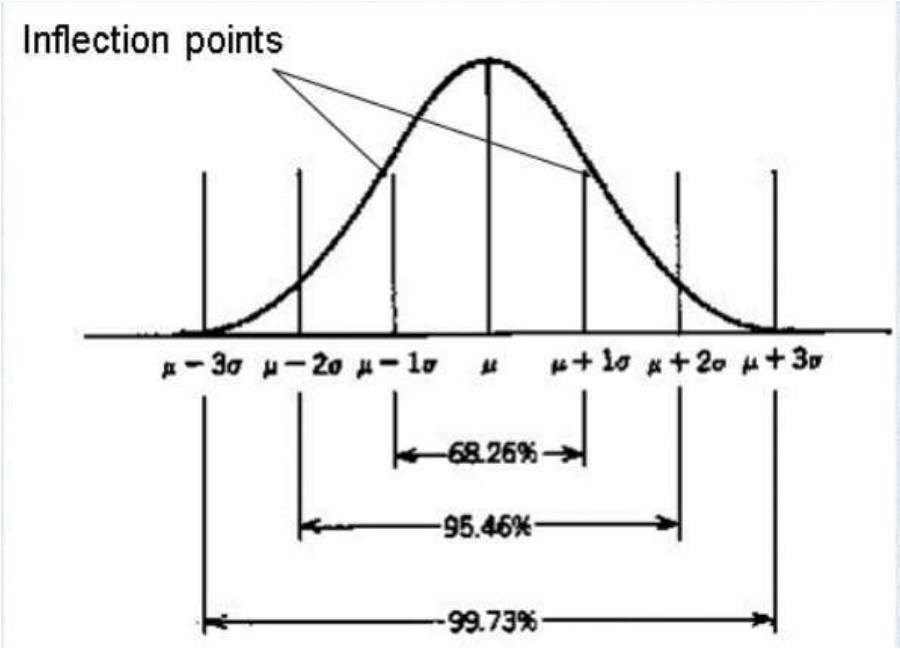
Exp 3b- Normal Distribution

In a random collection of data from independent sources, it is generally observed that the distribution of data is normal. Which means, on plotting a graph with the value of the variable in the horizontal axis and the count of the values in the vertical axis we get a bell shape curve. The centre of the curve represents the mean of the data set. In the graph, fifty percent of values lie to the left of the mean and the other fifty percent lie to the right of the graph. This is referred as normal distribution in statistics.

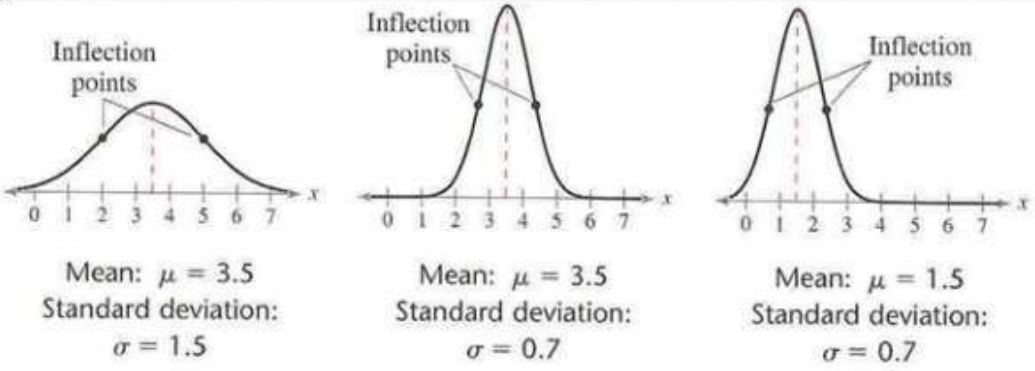
Properties:

A normal distribution is a continuous probability distribution for a random variable, x. The graph of a normal distribution is called the normal curve. A normal distribution has the following properties.

1. The mean, median and mode are equal.
2. The normal curve is bell-shaped and is symmetric about the mean.
3. The total area under the normal curve is equal to 1.
4. The normal curve approaches, but never touches the x-axis as it extends farther and farther away from the mean.
5. Between $\mu - \sigma$ and $\mu + \sigma$ (in the center of the curve) the graph curves downward. The graph curves upward to the left of $\mu - \sigma$ and to the right of $\mu + \sigma$. The points at which the curve changes from curving upward to curving downward are called **inflection points.**



6. A normal distribution can have any mean and any positive standard deviation. These two parameters, μ and σ completely determine the shape of a normal curve. The mean gives the location of the line of symmetry and the standard deviation describes how much the data are spread out.

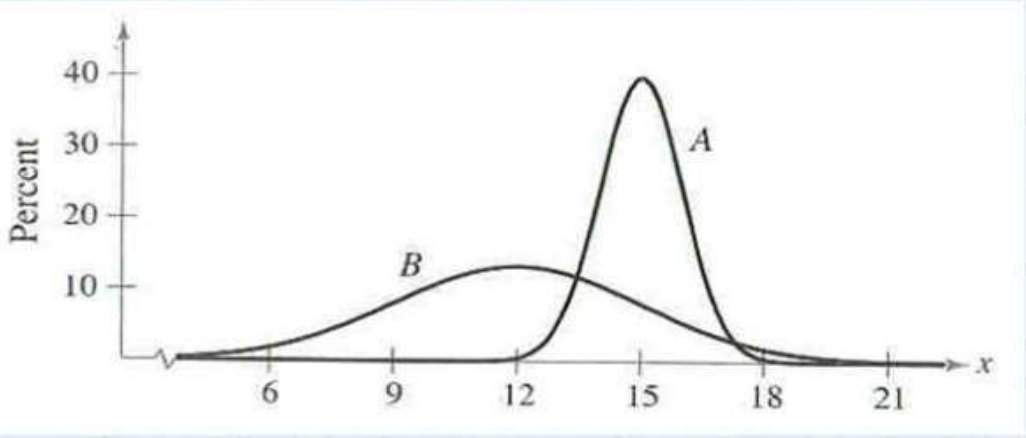


See the line of symmetry for each? That's the mean. However, if it is fatter, then the standard deviation is greater. That's the difference.

Understanding Mean & Standard Deviation

Which normal curve has a greater mean?

Which normal curve has a greater standard deviation



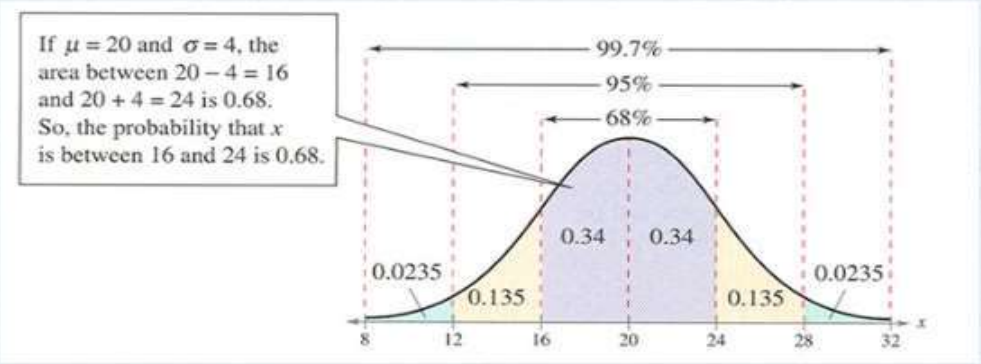
The line of symmetry of curve A occurs at x = 15. The line of symmetry of curve B occurs at x = 12. So, curve A has a greater mean.

Curve B is more spread out than curve A, so curve B has a greater standard deviation.

The Empirical Rule

In a normal distribution with mean μ and standard deviation σ, you can approximate areas under the normal curve as follows:

1. About 68% of the area lies between μ - σ and μ + σ
2. About 95% of the area lies between μ - 2σ and μ + 2σ
3. About 99.7% of the area lies between μ - 3σ and μ + 3σ



Normal Distribution

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

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

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 $Z = \frac{X-\mu}{\sigma}$ follows standard normal distribution (mean 0 and unit variance)

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

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 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)

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

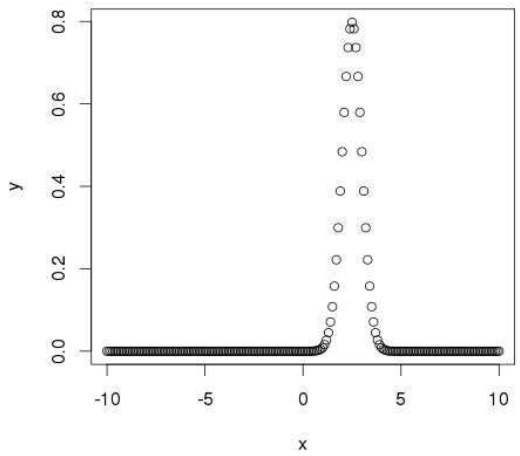
- **x** is a vector of numbers.
- **p** is a vector of probabilities.
- **n** is number of observations (sample size).
- **mean** is the mean value of the sample data. Its default value is zero.
- **sd** is the standard deviation. Its default value is 1.

dnorm()

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

```
# Create a sequence of numbers between -10 and 10 incrementing by 0.1.
x <- seq(-10, 10, by = .1)

# Choose the mean as 2.5 and standard deviation as 0.5.
y <- dnorm(x, mean = 2.5, sd = 0.5)

# Give the chart file a name.
png(file = "dnorm.png")
plot(x,y)

# Save the file.
dev.off()
```

When we execute the above code, it produces the following result −

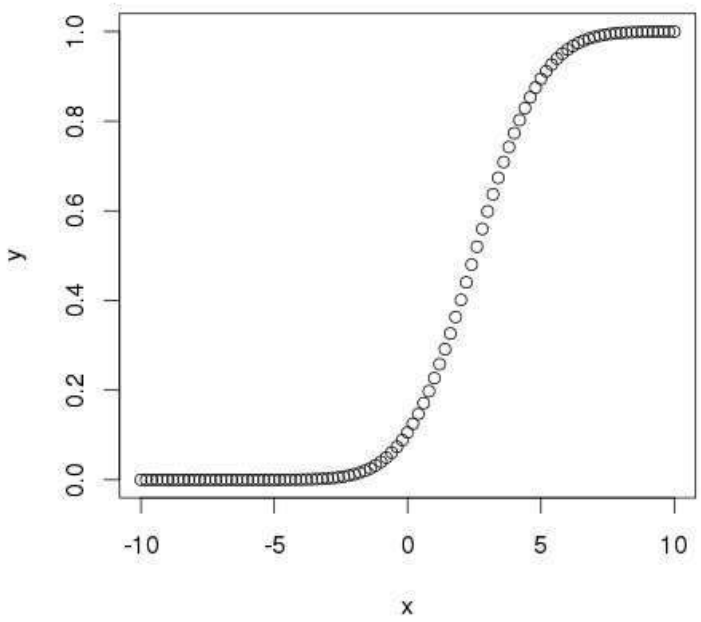


pnorm()

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".

```
# Create a sequence of numbers between -10 and 10 incrementing by 0.2.
x <- seq(-10,10,by = .2)

# Choose the mean as 2.5 and standard deviation as 2.
y <- pnorm(x, mean = 2.5, sd = 2)

# Give the chart file a name.
png(file = "pnorm.png")

# Plot the graph.
plot(x,y)

# Save the file.
dev.off()
```

When we execute the above code, it produces the following result −

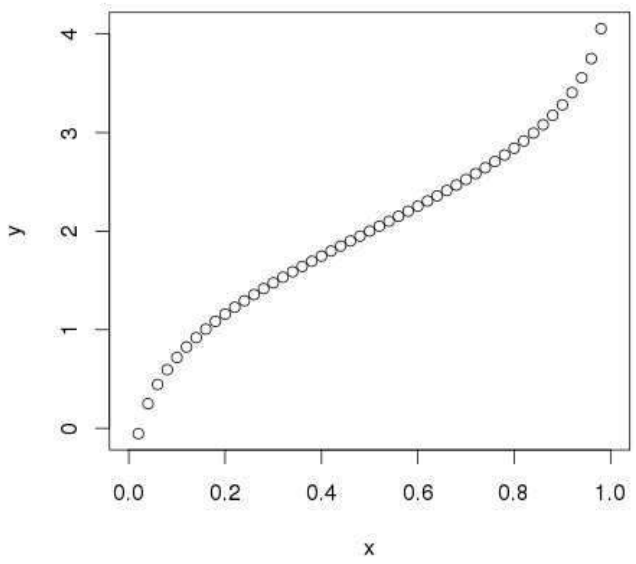


qnorm()

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

```
# Create a sequence of probability values incrementing by 0.02.
x <- seq(0, 1, by = 0.02)

# Choose the mean as 2 and standard deviation as 3.
y <- qnorm(x, mean = 2, sd = 1)

# Give the chart file a name.
png(file = "qnorm.png")

# Plot the graph.
plot(x,y)

# Save the file.
dev.off()
```

When we execute the above code, it produces the following result –

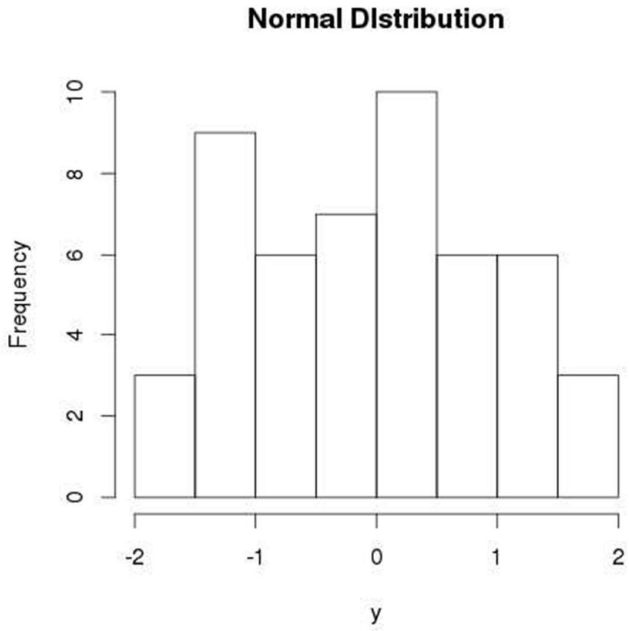


rnorm()

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.

```
# Create a sample of 50 numbers which are normally distributed.
y <- rnorm(50)

# Give the chart file a name.
png(file = "rnorm.png")

# Plot the histogram for this sample.
hist(y, main = "Normal DIstribution")

# Save the file.
dev.off()
```

When we execute the above code, it produces the following result −

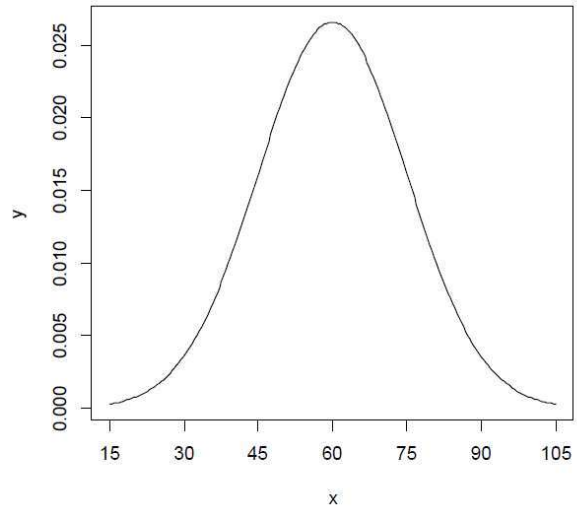


1. The weekly wages of 1000 workmen are normally distributed around a mean of Rs. 70 with S.D of Rs 5.Estimate the number of workers whose weekly wages will be
(i) Between Rs 69 and Rs 72
(ii) Less than Rs 69
(iii) More than Rs 72

```
> #(i)Between Rs 69 and Rs 72
> ( pnorm(72, mean=70, sd=5) - pnorm(69, mean=70, sd=5))*1000
[1] 234.6815
> #The number of workers whose wages lies between Rs.69 and Rs.72 is 234
> #(ii) Less than Rs 69
> (pnorm(69, mean=70, sd=5))*1000
[1] 420.7403
> #The number of workers whose wages is less than Rs.69  is 421
> #(iii) More than Rs 72
> (1 - pnorm(72, mean=70, sd=5))*1000
[1] 344.5783
> #The number of workers whose wages is More than Rs.72  is 345
```

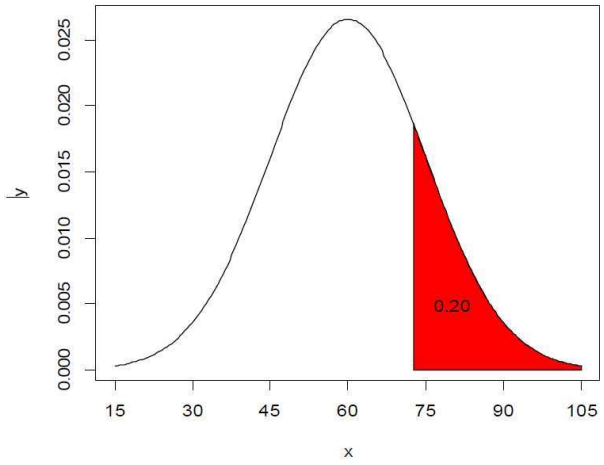
2. Draw a normal distribution with a mean=60 and a standard deviation=15.
>x=seq(15,105,length=200)
>y=dnorm(x,mean=60,sd=15)

```
>plot(x,y,type="l",xaxt="n")
>axis(1,at=c(15,30,45,60,75,90,105))
```



3.Shade the top 20% of the area under the normal density curve

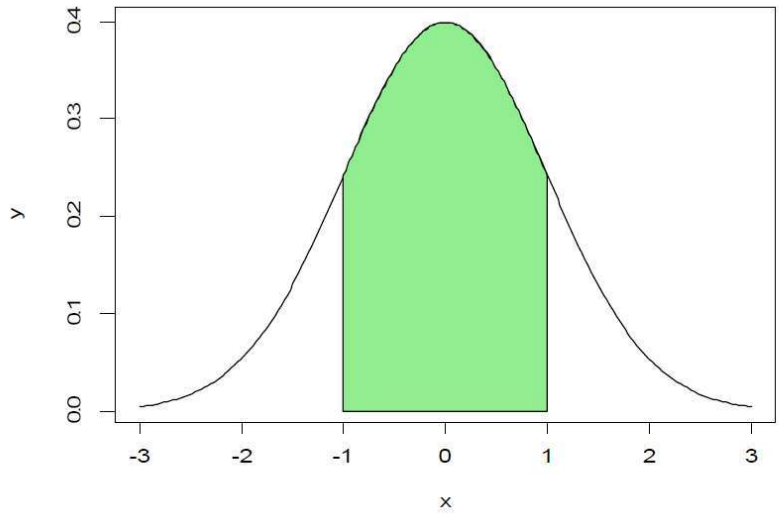
```
>x=seq(15,105,length=200)
>y=dnorm(x,mean=60,sd=15)
>plot(x,y,type="l",xaxt="n")
>axis(1,at=c(15,30,45,60,75,90,105))
>x=seq(72.62,105,length=100)
>y=dnorm(x,mean=60,sd=15)
>polygon(c(72.62,x,105),c(0,y,0),col="red")
>text(80,0.005,"0.20")
```



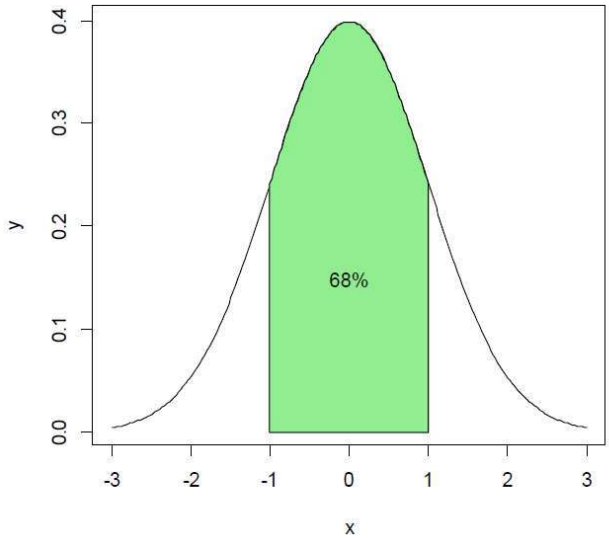
3. Simulate a standard normal density curve (mean=0 and standard deviation=1)

```
>x=seq(-3,3,length=200)
>y=dnorm(x)
```

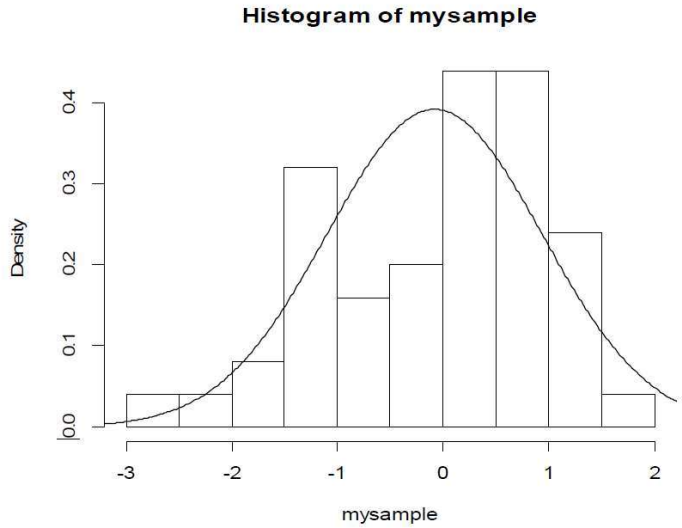
>plot(x,y,type="l")
>x=seq(-1,1,length=100)
>y=dnorm(x)
>polygon(c(-1,x,1),c(0,y,0),col="lightgreen")



>text(0,0.15,"68%")

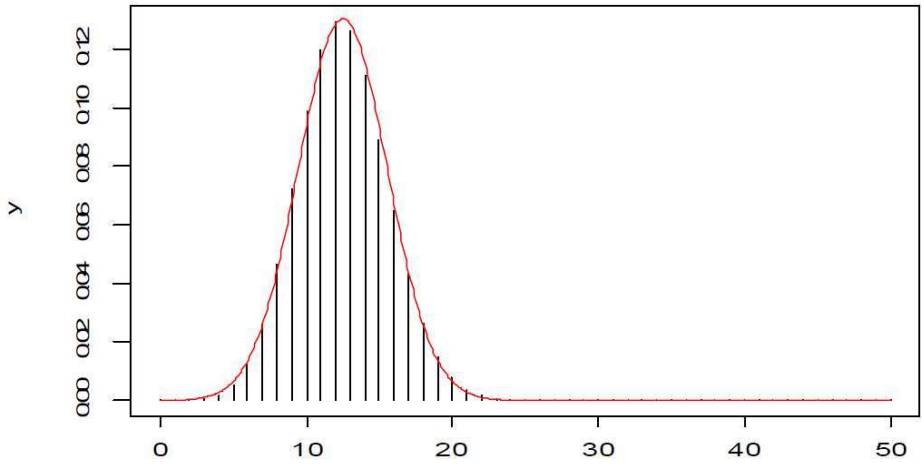


4. Generate 50 (standard) normally distributed random numbers and to display them as a histogram.
>mysample <- rnorm(50)
>hist(mysample, prob = TRUE)
>mu <- mean(mysample)
>sigma <- sd(mysample)
>x <- seq(-4, 4, length = 500)
>y <- dnorm(x, mu, sigma)
>lines(x,y)



5. Approximation of the binomial distribution with the normal distribution

```
> x <- 0:50
> y <- dbinom(x, 50, 0.25)
> plot(x, y, type="h")
> x2 <- seq(0, 50, length = 500)
> y2 <- dnorm(x2, 50*0.25, sqrt(50*0.25*(1-0.25)))
> lines(x2, y2, col = "red")
```



Practice:-

1. Suppose X is normal with mean 527 and
standard deviation 105. Compute P(X <=310)

```
>pnorm(310,527,105)
[1] 0.01938279
```

2. Find P(80 pts< x<95 pts.)

```
>pnorm(95, mean=100, sd=15) - pnorm(80,mean=100, sd=15)
[1] 0.2782301
```

3. In a test on 2000 Electric bulbs ,it was found that the life of particular make, was normally distributed with an average life of 2040 hours and S.D of 60 hours. Estimate the number of bulbs likely to burn for:
(i) More than 2150 hours
(ii) Less than 1950 hours
(iii) More than 1920 hours but less than 2160 hours
(iv) More than 2150 hours

```
> (1 - pnorm(2150, mean=2040, sd=60))*2000
[1] 66.75302
> (pnorm(1950, mean=2040, sd=60))*2000
[1] 133.6144
> ( pnorm(2160, mean=2040, sd=60)-pnorm(1920,mean=2040,sd=60))*2000
[1] 1908.999
```

(i) The number of bulbs expected to burn for more than 2150 hours is 67 (approximately)
(ii) The number of bulbs expected to burn for less than 1950 hours is 134 (approximately)
(iii) The number of bulbs expected to burn more than 1920 hours but less than 2160 is 1909 (approximately

References

Becker, R. A., Chambers, J. M. and Wilks, A. R. (1988) *The New S Language*. Wadsworth & Brooks/Cole.

Johnson, N. L., Kotz, S. and Balakrishnan, N. (1995) *Continuous Univariate Distributions*, volume 1, chapter 13. Wiley, New York.

Lab-8

One Sample Z-test

Aim:-To test the hypothesis for Large Samples by using one –sample Z-test

Procedure and R code:-

Test for significance of single mean:

Lower Tail Test of Population Mean with Known Variance:

The null hypothesis of the lower tail test of the population mean can be expressed as follows:

$$\mu \geq \mu_0$$

where μ_0 ***is a hypothesized lower bound of the true population mean*** μ ***.***
Let us define the test statistic z in terms of the sample mean, the sample size and thepopulation standard deviation σ ***:***

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} \sim N(0,1)$$

Then the null hypothesis of the lower tail test is to be rejected if $z \leq -z_\alpha$ ***, where*** z_α ***is the 100(1− α) percentile of the standard normal distribution.***

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 .05 significance level, can we reject the claim by the manufacturer?

R code:

The null hypothesis is that $\mu \geq 10000$. We begin with computing the test statistic.

```
> xbar = 9900              # sample mean
> mu0 = 10000              # hypothesized value
> sigma = 120              # population standard deviation
> n = 30                   # sample size
> z = (xbar-mu0)/(sigma/sqrt(n))
> z                        # test statistic
[1] -4.564355
```

We then compute the critical value at .05 significance level.

```
> alpha = .05
> z.alpha = qnorm(1-alpha)
> -z.alpha                 # critical value
[1] -1.644854
```

Interpretation:-

The test statistic -4.5644 is less than the critical value of -1.6449. Hence, at .05 significance level, we reject the claim that mean lifetime of a light bulb is above 10,000 hours.

Alternative Solution(compare with P value)

Instead of using the critical value, we apply the pnorm function to compute the lower tail p-value of the test statistic. As it turns out to be less than the .05 significance level, we reject the null hypothesis that $\mu \geq 10000$.

```
> pval = pnorm(z)
> pval                     # lower tail p-value
[1] 2.505166e-06
```

Upper Tail Test of Population Mean with Known Variance:

The null hypothesis of the upper tail test of the population mean can be expressed as follows:

$$\mu \geq \mu_0$$

where μ_0 is a hypothesized upper bound of the true population mean μ.
Let us define the test statistic z in terms of the sample mean, the sample size and thepopulation standard deviation σ :

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} \sim N(0,1)$$

Then the null hypothesis of the upper tail test is to be rejected if $z \geq z_\alpha$ ***where*** z_α ***is the 100(1− α) percentile of the standard normal distribution***

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 .05 significance level, can we reject the claim on food label?

R code:-

The null hypothesis is that $\mu \leq 2$***. We begin with computing the test statistic.***

```
> xbar = 2.1                  # sample mean
> mu0 = 2                     # hypothesized value
> sigma = 0.25                # population standard deviation
> n = 35                      # sample size
> z = (xbar-mu0)/(sigma/sqrt(n))
> z                           # test statistic
[1] 2.366432
```

We then compute the critical value at .05 significance level.

```
> alpha = .05
> z.alpha = qnorm(1-alpha)
> z.alpha                     # critical value
[1] 1.644854
```

Interpretation:-

The test statistic 2.3664 is greater than the critical value of 1.6449. Hence, at .05 significance level, we reject the claim that there is at most 2 grams of saturated fat in a cookie.

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 ***.Let us define the test statistic z in terms of the sample mean, the sample size and the population standard deviation*** σ ***:***

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} \sim N(0,1)$$

Then the null hypothesis of the two-tailed test is to be rejected if $z \leq -z_{\alpha/2}$ ***or*** $z \geq z_{\alpha/2}$ ***, where*** $z_{\alpha/2}$ ***is the*** $100(1 - \alpha/2)$ ***percentile of the standard normal distribution.***

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?

Solution

The null hypothesis is that μ = 15.4. We begin with computing the test statistic.

```
> xbar = 14.6              # sample mean
> mu0 = 15.4               # hypothesized value
> sigma = 2.5              # population standard deviation
> n = 35                   # sample size
> z = (xbar-mu0)/(sigma/sqrt(n))
> z                        # test statistic
[1] -1.893146
```

We then compute the critical values at .05 significance level.

```
> alpha = .05
> z.half.alpha = qnorm(1-alpha/2)
> c(-z.half.alpha, z.half.alpha)
[1] -1.959964  1.959964
```

Interpretation :

The test statistic -1.8931 lies between the critical values -1.9600 and 1.9600. Hence, at .05 significance level, we do not reject the null hypothesis that the mean penguin weight does not differ from last year.

Alternative Solution

Instead of using the critical value, we apply the pnorm function to compute the two-tailed p-value of the test statistic. It doubles the lower tail p-value as the sample mean is less than the hypothesized value. Since it turns out to be greater than the .05 significance level, we do not reject the null hypothesis that μ = 15.4.

```
> pval = 2 * pnorm(z)     # lower tail
> pval                    # two-tailed p-value
[1] 0.05833852
```

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.
Let us define the test statistic z in terms of the sample proportion and the sample size:

$$z = \frac{p - p_0}{\sqrt{p_0 q_0 / n}} \sim N(0,1)$$

Then the null hypothesis of the lower tail test is to be rejected if $z \leq -z_\alpha$, where z_α is the 100(1− α) percentile of the standard normal distribution.

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 0.5 significance level, can we reject the null hypothesis that the proportion of voters in the population is above 60% this year?

Solution

The null hypothesis is that $p \geq 0.6$. We begin with computing the test statistic.

```
> pbar = 85/148           # sample proportion
> p0 = .6                 # hypothesized value
> n = 148                 # sample size
> z = (pbar-p0)/sqrt(p0*(1-p0)/n)
> z                       # test statistic
[1] -0.6375983
```

We then compute the critical value at .05 significance level.

```
> alpha = .05
> z.alpha = qnorm(1-alpha)
> -z.alpha                # critical value
[1] -1.644854
```

Interpretation :

The test statistic -0.6376 is not less than the critical value of -1.6449. Hence, at .05 significance level, we do not reject the null hypothesis that the proportion of voters in the population is above 60% this year.

Alternative Solution 1

Instead of using the critical value, we apply the pnorm function to compute the lower tail p-value of the test statistic. As it turns out to be greater than the .05 significance level, we do not reject the null hypothesis that $p \geq 0.6$.

```
> pval = pnorm(z)
> pval
[1] 0.2618676
```

Alternative Solution 2

We apply the prop.test function to compute the p-value directly. The Yates continuity correction is disabled for pedagogical reasons.

Upper Tail Test of Population Proportion

The null hypothesis of the upper tail test about population proportion can be expressed as follows:

$$p \leq p_0$$

where p_0 is a hypothesized upper bound of the true population proportion p. Let us define the test statistic z in terms of the sample proportion and the sample size:

$$z = \frac{p - p_0}{\sqrt{p_0 q_0 / n}} \sim N(0,1)$$

Then the null hypothesis of the upper tail test is to be *rejected* if $z \geq z_\alpha$, where z_α is the $100(1-\alpha)$ percentile of the standard normal distribution.

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?

Solution

The null hypothesis is that $p \leq 0.12$. We begin with computing the test statistic.

```
> pbar = 30/214              # sample proportion
> p0 = .12                   # hypothesized value
> n = 214                    # sample size
> z = (pbar-p0)/sqrt(p0*(1-p0)/n)
> z                          # test statistic
[1] 0.908751
```

We then compute the critical value at .05 significance level.

```
> alpha = .05
> z.alpha = qnorm(1-alpha)
> z.alpha                    # critical value
[1] 1.644854
```

Interpretation:-

The test statistic 0.90875 is not greater than the critical value of 1.6449. Hence, at .05 significance level, we do not reject the null hypothesis that the proportion of rotten apples in harvest stays below 12% this year.

Alternative Solution 1

Instead of using the critical value, we apply the pnorm function to compute the upper tail p-value of the test statistic. As it turns out to be greater than the .05 significance level, we do not reject the null hypothesis that $p \leq 0.12$.

```
> pval = pnorm(z, lower.tail=FALSE)
> pval                    # upper tail p?value
[1] 0.1817408
```

Alternative Solution 2:

We apply the prop.test function to compute the p-value directly. The Yates continuity correction is disabled for pedagogical reasons.

```
> prop.test(30, 214, p=.12, alt="greater", correct=FALSE)

        1-sample proportions test without continuity correction

data:  30 out of 214, null probability 0.12
X-squared = 0.82583, df = 1, p-value = 0.1817
alternative hypothesis: true p is greater than 0.12
95 percent confidence interval:
 0.1056274 1.0000000
sample estimates:
        p
0.1401869
```

Two-Tailed Test of Population Proportion:

The null hypothesis of the two-tailed test about population proportion can be expressed as follows:

$$p = p_0$$

where p_0 is a hypothesized value of the true population proportion p.
Let us define the test statistic z in terms of the sample proportion and the sample size:

$$z = \frac{p - p_0}{\sqrt{p_0 q_0 / n}} \sim N(0,1)$$

Then the null hypothesis of the two-tailed test is to be *rejected* if $z \leq -z_{\alpha/2}$ or $z \geq z_{\alpha/2}$, where $z_{\alpha/2}$is the 100(1 − α) percentile of the standard normal distribution.

Problem

Suppose a coin toss turns up 12 heads out of 20 trials. At .05 significance level, can one reject the null hypothesis that the coin toss is fair?

Solution

The null hypothesis is that p = 0.5. We begin with computing the test statistic.

```
> pbar = 12/20              # sample proportion
> p0 = .5                   # hypothesized value
> n = 20                    # sample size
> z = (pbar-p0)/sqrt(p0*(1-p0)/n)
> z                         # test statistic
[1] 0.8944272
```

We then compute the critical values at .05 significance level.

```
> alpha = .05
> z.half.alpha = qnorm(1-alpha/2)
> c(-z.half.alpha, z.half.alpha)
[1] -1.959964  1.959964
```

Interpretation:

The test statistic 0.89443 lies between the critical values -1.9600 and 1.9600. Hence, at .05 significance level, we do not reject the null hypothesis that the coin toss is fair.

Alternative Solution 1

Instead of using the critical value, we apply the pnorm function to compute the two-tailed p-value of the test statistic. It doubles the upper tail p-value as the sample proportion is greaterthan the hypothesized value. Since it turns out to be greater than the .05 significance level, we do not reject the null hypothesis that p = 0.5.

```
>  pval = 2 * pnorm(z, lower.tail=FALSE)  # upper tail
>  pval
[1] 0.3710934
```

Alternative Solution 2

We apply the prop.test function to compute the p-value directly. The Yates continuity correction is disabled for pedagogical reasons.

```
> prop.test(12, 20, p=0.5, correct=FALSE)

        1-sample proportions test without continuity correction

data:  12 out of 20, null probability 0.5
X-squared = 0.8, df = 1, p-value = 0.3711
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.3865815 0.7811935
sample estimates:
  p
0.6
```

Practice Problems:

1. **A sample of 100 tyres is taken from a lot. The mean life of tyres is found to be 39, 350 kilo meters with a standard deviation of 3, 260. Could the sample come from a population with mean life of 40, 000 kilometers?**

2. **the mean life time of a sample of 400 fluorescent light bulbs produced by a company is found to be 1, 570 hours with a standard deviation of 150 hours. Test the hypothesis that the mean life time of bulbs is 1600 hours against the alternative hypothesis that it is greater than 1, 600 hours at 1% and 5% level of significance.**
3. **In the sample of 1000 people in Maharashtra,540 are rice eaters and the rest are wheat eaters. Can we assume that both rice and wheat are equally popular in this state at 1% level of significance**
4. **A particular brand of tires claims that its deluxe tire averages at least 50,000 miles before it needs to be replaced. From past studies of this tire, the standard deviation is known to be 8000. A survey of owners of that tire design is conducted. From the 28 tires surveyed, the average lifespan was 46,500 miles with a standard deviation of 9800 miles. Do the data support the claim at the 5% level?**
5. **From generation to generation, the average age when smokers first start to smoke varies. However, the standard deviation of that age remains constant of around 2.1 years. A survey of 40 smokers of this generation was done to see if the average starting age is at least 19. The sample average was 18.1 with a sample standard deviation of 1.3. Do the data support the claim at the 5% level?**
6. **The cost of a daily newspaper varies from city to city. However, the variation among prices remains steady with a standard deviation of 6¢. A study was done to test the claim that the average cost of a daily newspaper is 35¢. Twelve costs yield an average cost of 30¢ with a standard deviation of 4¢. Do the data support the claim at the 1% level?**
7. **An article in the *San Jose Mercury News* stated that students in the California state university system take an average of 4.5 years to finish their undergraduate degrees. Suppose you believe that the average time is longer. You conduct a survey of 49 students and obtain a sample mean of 5.1 with a sample standard deviation of 1.2. Do the data support your claim at the 1% level?**
8. **The average number of sick days an employee takes per year is believed to be about 10. Members of a personnel department do not believe this figure. They randomly survey 8 employees. The number of sick days they took for the past year are as follows: 12; 4; 15; 3; 11; 8; 6; 8. Let X = the number of sick days they took for the past year. Should the personnel team believe that the average number is about 10?**
9. **In 1955, *Life Magazine* reported that the 25 year-old mother of three worked [on average] an 80 hour week. Recently, many groups have been studying whether or not the women's movement has, in fact, resulted in an increase in the average work week for women (combining employment and at-home work). Suppose a study was done to determine if the average work week has increased. 81 women were surveyed with the following results. The sample average was 83; the sample standard deviation was 10. Does it appear that the average work week has increased for women at the 5% level?**

10. **Your statistics instructor claims that 60 percent of the students who take her Elementary Statistics class go through life feeling more enriched. For some reason that she can't quite figure out, most people don't believe her. You decide to check this out on your own. You randomly survey 64 of her past Elementary Statistics students and find that 34 feel more enriched as a result of her class. Now, what do you think?**

11. **According to an article in *Newsweek*, the natural ratio of girls to boys is 100:105. In China, the birth ratio is 100: 114 (46.7% girls). Suppose you don't believe the reported figures of the percent of girls born in China. You conduct a study. In this study, you count the number of girls and boys born in 150 randomly chosen recent births. There are 60 girls and 90 boys born of the 150. Based on your study, do you believe that the percent of girls born in China is 46.7?**
12. **A poll done for *Newsweek* found that 13% of Americans have seen or sensed the presence of an angel. A contingent doubts that the percent is really that high. It conducts its own survey. Out of 76 Americans surveyed, only 2 had seen or sensed the presence of an angel. As a result of the contingent's survey, would you agree with the *Newsweek* poll? In complete sentences, also give three reasons why the two polls might give different results.**

LAB-11

Paired t-test And F- (Variance Ratio Test)

AIM: to analyse the improvement or effectiveness of a new methodology adopted. And also test the hypothesis for variance ratio.

HYPOTHESIS TESTS FOR MEAN DIFFERENCES: PAIRED DATA-t-TEST

Problem 1 :

A school athletics has taken a new instructor, and want to test the effectiveness of the new type of training proposed by the new instructor comparing the average times of 10 runners in the 100 meters. The results are given below(time in seconds)

Before training	*12.9*	*13.5*	*12.8*	*15.6*	*17.2*	*19.2*	*12.6*	*15.3*	*14.4*	*11.3*
After training	*12.7*	*13.6*	*12.0*	*15.2*	*16.8*	*20.0*	*12.0*	*15.9*	*16.0*	*11.1*

Solution: *In this case we have two sets of paired samples, since the measurements were made on the same athletes before and after the workout. To see if there was an improvement, deterioration, or if the means of times have remained substantially the same (hypothesis H0), we need to make a Student's t-test for paired samples, proceeding in this way*

```
> before = c(12.9, 13.5, 12.8, 15.6, 17.2, 19.2, 12.6, 15.3, 14.4, 11.3)
> after = c(12.7, 13.6, 12.0, 15.2, 16.8, 20.0, 12.0, 15.9, 16.0, 11.1)
> t.test(before,after, paired=TRUE)

        Paired t-test

data:  before and after
t = -0.21331, df = 9, p-value = 0.8358
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.5802549  0.4802549
sample estimates:
mean of the differences
                  -0.05
```

Interpretation :-

The p-value is greater than 0.05, then we do not reject the hypothesis H_0 of equality of the averages and conclude that the new training has not made any significant improvement to the team of athletes.

Problem 2 :-

Suppose now that the manager of the team (given the results obtained) fired the coach who has not made any improvement, and take another, more promising. We report the times of athletes after the second training:

Before training:	*12.9*	*13.5*	*12.8*	*15.6*	*17.2*	*19.2*	*12.6*	*15.3*	*14.4*	*11.3*
After the second training:	*12.0*	*12.2*	*11.2*	*13.0*	*15.0*	*15.8*	*12.2*	*13.4*	*12.9*	*11.0*

R code:-

Now we check if there was actually an improvement, ie perform a t-test for paired data, specifying in R to test the alternative hypothesis H1 of improvement in times. To do this simply add the syntax alt = "less" when you call the t-test:

```
> before=c(12.9, 13.5, 12.8, 15.6, 17.2, 19.2, 12.6, 15.3, 14.4, 11.3)
> after = c(12.0, 12.2, 11.2, 13.0, 15.0, 15.8, 12.2, 13.4, 12.9, 11.0)
> t.test(before,after, paired=TRUE, alt="less")

        Paired t-test

data:  before and after
t = 5.2671, df = 9, p-value = 0.9997
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
     -Inf 2.170325
sample estimates:
mean of the differences
                  1.61
```

Interpretation:-

In response, we obtained a p-value well above 0.05, which leads us to conclude that we can reject the null hypothesis H_0 in favour of the alternative hypothesis H1: the new training has made substantial improvements to the team.

Problem 3 *: Consider the paired data below that represents cholesterol levels on 10 men before and after a certain medication*

Before(x)	*237*	*289*	*257*	*228*	*303*	*275*	*262*	*304*	*244*	*233*
After(y)	*194*	*240*	*230*	*186*	*265*	*222*	*242*	*281*	*240*	*212*

Test the claim that, on average, the drug lowers cholesterol in all men. I.e., test the claim that $\mu_d > 0$. *Test this at the 0.05 significance level.*

R-code:-

```
> before=c(237,289,257,228,303,275,262,304,244,233)
> after=c(194,240,230,186,265,222,242,281,240,212)
> t.test(before,after,paired=TRUE,alternative="greater",mu=0)

        Paired t-test

data:  before and after
t = 6.5594, df = 9, p-value = 5.202e-05
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 23.05711      Inf
sample estimates:
mean of the differences
                     32
```

Interpretation :-

We can reject the null hypothesis and support the claim because the P-value ($\approx 5.2X10^{-5}$) is less than the signicance level.

F Test to Compare Two Variances

Problem 1 :-

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*

R code:

$H_0: S_1^2 = S_2^2$
$H_{1:} S_1^2 \neq S_2^2$
Level of Significance :0.10

R code:-

```
> Unit_A=c(14.1,10.1,14.7,13.7,14.0)
> Unit_B=c(14.0,14.5,13.7,12.7,14.1)
> var.test(Unit_A,Unit_B)

        F test to compare two variances

data:  Unit_A and Unit_B
F = 7.3304, num df = 4, denom df = 4, p-value = 0.07954
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
  0.7632268 70.4053799
sample estimates:
ratio of variances
          7.330435
```

Inference : here p valuve >0.05 ,then there is no evidence to reject the null hypothesis.

Problem 2: Energy Data :- (Variance Ratio-test)

```
> energy=read.csv("C:\\Users\\aadmin\\Desktop\\energy.csv")
> energy
   expend stature
1    9.21   obese
2    7.53    lean
3    7.48    lean
4    8.08    lean
5    8.09    lean
6   10.15    lean
7    8.40    lean
8    0.88    lean
9    6.13    lean
10   7.90    lean
11  11.51   obese
12  12.79   obese
13   7.05    lean
14  11.85   obese
15   9.97   obese
16   7.48    lean
17   8.79   obese
18   9.69   obese
19   9.68   obese
20   7.58    lean
21   9.19   obese
22   8.11    lean
> var.test(energy$expend~energy$stature)

        F test to compare two variances

data:  energy$expend by energy$stature
F = 2.321, num df = 12, denom df = 8, p-value = 0.2386
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.5526712 8.1509583
sample estimates:
ratio of variances
  |        2.321035
```

Inference :

Here p valuve >0.05 ,then there is no evidence to reject the null hypothesis.

Practice questions:-

1. *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 mileage for that tank was recorded. The mileage was recorded again for the same car using the other kind of gasoline. We use a paired t – test to determine whether cars get significant better mileage with premium gas.*

Regular	***16***	***20***	***21***	***22***	***23***	***22***	***27***	***25***	***27***	***28***
Premium	*19*	*22*	*24*	*24*	*25*	*25*	*26*	*26*	*28*	

2. *The Scores of 10 candidates prior and after training are given below*

Prior	*84*	*48*	*36*	*37*	*54*	*69*	*83*	*96*	*90*	*65*
After	*90*	*58*	*56*	*49*	*62*	*81*	*84*	*86*	*84*	*75*

Test the training is Effective or Not?

3. *An IQ test was administrated to 5 persons before and after they were trained. The results are given below*

Candidates	*I*	*II*	*III*	*IV*	*V*
IQ before Training	*110*	*120*	*123*	*132*	*125*
IQ After Training	*120*	*118*	*125*	*136*	*121*

Test whether there is any change in IQ after the training Programme

4. *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

5. 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

Chi-square Test

Goodness of Fit and Independence of Attributes

Chi-square test for independence of attributes

Two random variables x and y are called independent if the probability distribution of one variable is not affected by the presence of another. Assume Oij is the observed frequency count of events belonging to both i-th category of x and j-th category of y. Also assume Eij to be the corresponding expected count if x and y are independent. The null hypothesis of the independence assumption is to be rejected if the p-value of the following Chi-squared test statistics is less than a given significance level α.

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

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*

R- Code and Interpretation

```
> data<-matrix(c(35,42,61,48,51,68),ncol=3,byrow=T)
> data
     [,1] [,2] [,3]
[1,]   35   42   61
[2,]   48   51   68
> chisq.test(data)

        Pearson's Chi-squared test

data:  data
X-squared = 0.53441, df = 2, p-value = 0.7655
```

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.

Problem

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)
> tbl = table(survey$Smoke, survey$Exer)
> tbl
```

	Freq	None	Some
Heavy	7	1	3
Never	87	18	84
Occas	12	3	4
Regul	9	1	7

Test the hypothesis whether the students smoking habit is independent of their exercise level at .05 significance level.

R- Code and Interpretation

```
> chisq.test(tbl)

        Pearson's Chi-squared test

data:  tbl
X-squared = 5.4885, df = 6, p-value = 0.4828

Warning message:
In chisq.test(tbl) : Chi-squared approximation may be incorrect
```

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.

```
> ctbl = cbind(tbl[,"Freq"], tbl[,"None"] + tbl[,"Some"])
> ct

> ctbl
      [,1] [,2]
Heavy    7    4
Never   87  102
Occas   12    7
Regul    9    8
```

R- Code

```
> chisq.test(ctbl)

        Pearson's Chi-squared test

data:  ctbl
X-squared = 3.2328, df = 3, p-value = 0.3571
```

Goodness of Fit

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?

R- Code & Inference

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

        Chi-squared test for given probabilities

data:  plants
X-squared = 1.9702, df = 3, p-value = 0.5786

Warning message:
In chisq.test(plants, p = c(9/16, 3/16, 3/16, 1/16)) :
  Chi-squared approximation may be incorrect
```

The Chi-squared distribution is only an approximation to the sampling distribution of our test statistic, and the approximation is not very good when the expected cell counts are too small. This is the reason for the warning.

Here the probability value p is greater than alpha level (0.05), so we do not reject the null hypothesis.

Fitting of Binomial Distribution with Goodness of Fit

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?

Solution :

Let us setup the null hypothesis that the data are consistent with the hypothesis of equal probability for male and female births.

R- CODE & INFERENCE

```
> x=c(5,4,3,2,1,0)
> n=5                          #Probability of 'r' male births in a family
> N=320                        #Total Number of families
> P<-0.5                       #Probability of Male Birth
> Obf<-c(14,56,110,88,40,12)   #Obsevred frequencies
> exf<-dbinom(x,n,P)*320       #Expected frequencies
> # check the Condintion Sum of Observed and Expected are Equal
> sum(Obf)
[1] 320
> sum(exf)
[1] 320
> chisq<-sum((Obf-exf)^2/exf)
> chisq
[1] 7.16
> qchisq(0.95,5)
[1] 11.0705
```

Calculated value of chi-square is less than the tabulated value ,it is not significant at 5 % level of significance and hence the null hypothesis of equal probability for male and female births.

Fitting of Poisson Distribution with Goodness of Fit

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)    #expcted frequencies
> f1=round(expf)
> # check obserevd and Expected frequencies Total
> sum(f)
[1] 392
> sum(f1)
[1] 393
> # here subtrat '1' from expected frequencies
> #The last 3 frequencies are less than 5 so combine these frequencies in Observation and Expected
> obf<-c(275,72,30,15)
> exf<-c(242,117,28,6)
> chisq<-sum(((obf-exf)^2)/exf)
> chisq
[1] 35.45055
> qchisq(0.95,2)
[1] 5.991465
```

Inference

Since calculated value of χ^2 =35.45055 is much greater than 5.99,it is highly significant .Hence we conclude that poisson distribution is not good fit to the given data

Fitting the Normal Distribution with Goodness of Fit

Problem : The following table displays a frequency distribution of heights of trees in a certain locality. Fit a normal distribution to the data and test the goodness of fit.

Class Interval	Frequency
13.20 – 20.90	2
20.90 – 28.60	10
28.60 – 36.30	16
36.30 – 44.00	37
44.00 – 51.70	43
51.70 – 59.40	39
59.40 – 67.10	29
67.10 – 74.80	13
74.80 – 82.50	06
82.50 – 90.20	05

Heights of Trees (in inches)

```
> midy<-seq(17.05,86.5,length=10)
> f<-c(2,10,16,37,43,39,29,13,6,5)
> mean<-sum(f*midy)/sum(f)
> sd<-sqrt(sum(f*(midy-mean)^2)/sum(f))
> l<-seq(13.2,82.5,length=10)
> l<-c(l,90.2)
> cdf<-pnorm(l,mean,sd)
> cdf<-c(0,cdf,1)
> pcf<-diff(cdf)
> f<-c(0,f,0)
> ex<-round(pcf*sum(f),4)
> fr<-data.frame(f,ex)
> obf<-c(12,16,37,43,39,29,13,11)
> exf<-c(sum(ex[c(1,2,3)]),ex[c(4:9)],sum(ex[c(10,11,12)]))
> sum(obf)
[1] 200
> sum(exf)
[1] 200
> chisq<-sum((obf-exf)^2/exf)
> chisq
[1] 2.153974
> qchisq(0.95,5)
[1] 11.0705
```

Inference

Here chi-square cal value is less than chi-square tab value then there is no evidence to reject our null hypothesis.ie the fit of normal distribution is good

Practice Problems

1. The following data come from a hypothetical survey of 920 people (Men, Women) that ask for their preference of one of the three ice cream flavors (Chocolate, Vanilla, Strawberry). Is there any association between gender and preference for ice cream flavor?

Gender\flavor	Chocolate	Vanilla	Strawberry
Men	100	120	60
Women	350	320	150

2. As a part of quality improvement project focused on a delivery of mail at a department office within a large company, data were gathered on the number of different addresses that had to be changed so that the mail could be redirected to thee correct mail stop. Table shows the frequency distribution. Fit binomial distribution and test goodness of fit

x	0	1	2	3	4
fx	5	20	45	20	10

The number of Addresses Needing Change

LAB-13

(Design of Experiments –CRD, RBD and LSD)

Completely Randomized Design (in Design of Experiments):

The conducting of an experiment by allotting treatments (factors), whose effects are to be experimented, to uniform/homogeneous experimental units by a simple random sampling design such that every unit can receive any treatment with equal chance, analyzing (splitting) total variation in the results into variation due to treatments and variation due to chance (error or residual) and then testing the significance or otherwise of treatments variation over error variation.

CRD or Allotment of Treatments in CRD: It can be explained easily with 4 treatments denoted by the letters A, B, C, D to be allotted to 16 experimental units of uniform quality.

Let A be repeated on 3 units; B on 6 units; C on 3 units; D on 4 units.

Random allotment (randomization) is done like this-----serialize the 16 units from 1 to 16 first; take 16 slips of paper of equal size and shape, write the letters A on 3 slips. B on 6 slips, C on 3 slips and D on 4 slips; fold/roll these slips well separately, put them in a box/bag, shuffle them well and take one slip after another; allot the first chosen letter (slip) to the first unit, 2^{nd} chosen letter to 2^{nd} unit and so on until all the letters (treatments) from the chosen slips are allotted to all the 16 units completely.

NOTE: *Randomization of letters can also be done by ' Random Number Tables' or by 'computerized randomization'.*

Then the resulting layout may be as shown below:

CRD- layout in square form.

B	*D*	*A*	*C*
A	*C*	*B*	*B*
D	*B*	*C*	*A*
B	*D*	*B*	*D*

CRD-layout may be in rectangular form also as shown below: Let us consider treatment A repeated 4 times, B repeated 2 times, C repeated 5 times and D repeated 4 times on 15 units.

CRD ANOVA TABLE:-

Source of variation	df	SS	Mean SS= MSS = $\frac{SS}{df}$	F- ratio
Treat-ments	r−1	$\sum_{i=1}^{r}\left(\frac{R_i^2}{n_i}\right)-CF$	$\left(\frac{Treatment\ SS}{r-1}\right)$	$\left(\frac{Treatt.MSS}{Error\ MSS}\right)\rightarrow F_{[(r-1),\ (\sum n_i - r)]}$
Error	$\sum_{i=1}^{r}(n_i)-r$	by subtraction	$\left\{\frac{Error\ SS}{\sum_{i=1}^{r}(n_i)-r}\right\}$	------
Total	$\sum_{i=1}^{r}(n_i)-1$	$\sum_{i=1}^{r}\sum_{j=1}^{n_i}x_{ij}^2-CF$	-----	

The F-ratio $=\left(\frac{\text{Treatt.MSS}}{\text{ErrorMSS}}\right)\rightarrow F_{\left(\overline{r-1,\ \sum n_i - r}\right)}$- distribution with (r–1) as the numerator degrees of freedom (df) and $[(\sum_{i=1}^{r} n_i)-r]$ as the denominator df.

In this table, Correction Factor, CF = $\left[\frac{(\text{Grand Total})^2}{\text{Total no.of observations}}\right]$.

Completle Randomised design (Method of ANOVA for one way-classificantion with equal number of Observations)

Problem: 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 average sales volume for the 3 new menu items are all equal.

Item 1	Item2	Item3
22	52	16
42	33	24
44	8	19
52	47	18
45	43	34
37	32	39

(Enter the Above data in Excel Sheet)

R code:-

```
> df1=read.csv("C:\\Users\\aadmin\\Desktop\\CRD.csv")
> df1
  Item.1 Item2 Item3
1     22     52     16
2     42     33     24
3     44      8     19
4     52     47     18
5     45     43     34
6     37     32     39
> r = c(t(as.matrix(df1)))   #resopnse data
> r
 [1] 22 52 16 42 33 24 44  8 19 52 47 18 45 43 34 37 32 39
> f = c("Item1", "Item2", "Item3")    #treatment levels
> f
[1] "Item1" "Item2" "Item3"
> k = 3      # number of treatment levels
> n = 6      # observations per treatment
> tm = gl(k, 1, n*k, factor(f)) # matching treatments
> tm
 [1] Item1 Item2 Item3 Item1 Item2 Item3 Item1 Item2 Item3 Item1 Item2 Item3
[13] Item1 Item2 Item3 Item1 Item2 Item3
Levels: Item1 Item2 Item3
> crdfit = aov(r ~ tm)
> summary(crdfit)
            Df Sum Sq Mean Sq F value Pr(>F)
tm           2  745.4   372.7   2.541  0.112
Residuals   15 2200.2   146.7
```

Interpretation : *Since the p-value of 0.112 is greater than the .05 significance level, we do not reject the null hypothesis that the mean sales volume of the new menu items are all equal*.

Completle Randomised design (Method of ANOVA for one way-classificantion with un equal number of Observations)

Problem :The following Table shows the lives(in hours) of four batches of electric lamps

Batches	Life of Bulbs in Hrs							
1	1600	1610	1650	1680	1700	1720	1800	
2	1580	1640	1640	1700	1750			
3	1460	1550	1600	1620	1640	1660	1740	1820
4	1510	1520	1530	1570	1600	1680		

(Enter the Above data in Excel Sheet)

R code:-

```
>data=c(1600,1610,1650,1680,1700,1720,1800,1580,1640,1640,1700,1750,1460,155
0,1600,1620,1640,1660,1740,1820,1510,1520,1530,1570,1600,1680)

>batchs=c("batch1","batch1","batch1","batch1","batch1","batch1","batch1","batch2",
"batch2","batch2","batch2","batch2","batch3","batch3","batch3","batch3","batch3","b
atch3","batch3","batch3","batch4","batch4","batch4","batch4","batch4","batch4")

> Anova1=aov(data~batchs)

> summary(Anova1)

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

batchs       3  44361   14787   2.149  0.123

Residuals   22 151351    6880
```

Interpretation :

We may regard the four batches of electric lamps to be homogeneous

Randomized Block Design (in Design of Experiments): The conducting of an experiment on experimental units, which differ in quality with respect to one character and hence stratified with respect to such changing character into different within-homogeneous blocks/strata and then allotting treatments, whose effects are to be experimented, to homogenized units within each block by simple random sampling design independently without repetitions and thereafter splitting the total variation in the results into blocks variation, treatments variation, residual /error variation and lastly testing the significance of these variations over error variation.

In symbols,

allotting treatments to experimental units, differing in quality with respect to one character, by stratified random sampling, by which

Total variation = Blocks variation + Treatments variation + Error variation

and then testing if Blocks variation and Treatments variation are significant over Error variation, obtained from the results on such randomized block design(RBD) data .

A specimen of RBD, with four treatments A,B,C,D in 5 blocks taken in rows, say is given below:

with rows as blocks---→

Block-I	D	C	A	B
Block-II	B	A	C	D
Block-III	A	D	B	C
Block-IV	C	D	B	A
Block-V	B	A	D	C

Or

With columns as blocks---→

Block				
I	II	III	IV	V
C	A	B	D	A
B	C	A	C	D
D	B	C	A	B
A	D	D	B	C

R.B.D ANOVA table:-

Source of Variation	**df**	**SS**	**Mean SS= $\frac{SS}{df}$**	**F-ratio (F-distribution)**
Rows (Blocks)	r – 1	$\frac{\sum_{i=1}^{r} R_i^2}{c} - CF$	$\frac{\text{Row SS}}{r-1}$	$\left(\frac{\text{Row MSS}}{\text{Error MSS}}\right) \rightarrow F_{[(r-1),\ (r-1)(c-1)]}$
Columns (Treatm-ents)	c – 1	$\frac{\sum_{j=1}^{c} C_j^2}{r} - CF$	$\frac{\text{Column SS}}{c-1}$	$\left(\frac{\text{Column MSS}}{\text{Error MSS}}\right) \rightarrow F_{[(c-1),\ (r-1)(c-1)]}$
Error	(r –1) (c –1)	(by subtra-ction)	$\frac{\text{Error SS}}{(r-1)(c-1)}$	-----
Total	rc – 1	$\sum_{i=1}^{r}\sum_{j=1}^{c} {x_{ij}}^2 - CF$	-----	

In the table, MSS = Mean SS, CF = Correction factor = $\left[\frac{(GT)^2}{\text{Total no.of observations, } rc}\right]$ and df = degrees of freedom.

The F-ratio = $\left(\frac{\text{Row MSS}}{\text{Error MSS}}\right) \rightarrow F_{[(r-1),\ (r-1)(c-1)]}$ –distribution with (r – 1) as the numerator df and (r – 1)(c – 1) as the denominator (Error) df.

Similarly, the F-ratio = $\left(\frac{\text{Column MSS}}{\text{Error MSS}}\right) \rightarrow F_{[(c-1),\ (r-1)(c-1)]}$ –distribution with (c –1) as the numerator df and (r – 1)(c – 1) as the denominator (Error) df.

Two-way analysis of variance:

*Problem: Suppose each row in the following table represents the sales figures of the 3 new menu in a restaurant after a week of test marketing. At .05 level of significance, test whether the average sales volume for the 3 new menu items are all equal***.**

Data :-

ITEM1	ITEM2	ITEM3
31	27	24
31	28	31
45	29	46
21	18	48
42	36	46
32	17	40

```
> df2=read.csv("C:\\Users\\aadmin\\Desktop\\RBD.csv")
> df2
  ITEM1 ITEM2 ITEM3
1    31    27    24
2    31    28    31
3    45    29    46
4    21    18    48
5    42    36    46
6    32    17    40
> r = c(t(as.matrix(df2))) # response data
> r
 [1] 31 27 24 31 28 31 45 29 46 21 18 48 42 36 46 32 17 40
> f = c("Item1", "Item2", "Item3") # treatment levels
> k = 3 # number of treatment levels
> n = 6 # number of control blocks
> tm = gl(k, 1, n*k, factor(f)) # matching treatment
> tm
 [1] Item1 Item2 Item3 Item1 Item2 Item3 Item1 Item2 Item3 Item1 Item2 Item3
[13] Item1 Item2 Item3 Item1 Item2 Item3
Levels: Item1 Item2 Item3
> blk = gl(n, k, k*n) # blocking factor
> blk
 [1] 1 1 1 2 2 2 3 3 3 4 4 4 5 5 5 6 6 6
Levels: 1 2 3 4 5 6
> rbdfit = aov(r ~ tm + blk)
> summary(rbdfit) # Print out the ANOVA table with the summary function.
            Df Sum Sq Mean Sq F value Pr(>F)
tm           2  538.8  269.39   4.959 0.0319 *
blk          5  559.8  111.96   2.061 0.1547
Residuals   10  543.2   54.32
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interpretation :-

Since the p-value of 0.032 is less than the .05 significance level, we reject the null hypothesis that the mean sales volume of the new menu items are all equal.

Problem 2 : The data recorded for yield in a randomized block design experiment involving six treatments in four randomized blocks are given below. Analyse and interpret the data

	Treatments and yield					
Blocks	(1) 24.7	(3) 27.7	(2) 20.6	(4) 16.2	(5) 16.2	(6) 24.9
	(3) 22.7	(2) 28.8	(1) 27.3	(4) 15.0	(6) 22.5	(5) 17.0
	(6) 26.3	(4) 19.6	(1) 38.5	(3) 36.8	(2) 39.5	(5) 15.4
	(5) 17.7	(2) 31.0	(1) 28.5	(4) 14.1	(3) 34.9	(6) 22.6

Solution : Arrange this data in order

		Treatments and yield					
		1	**2**	**3**	**4**	**5**	**6**
Blocks	1	24.7	20.6	27.7	16.2	16.2	24.9
	2	27.3	28.8	22.7	15.0	17.0	22.5
	3	38.5	39.5	36.8	19.6	15.4	26.3
	4	28.5	31.0	34.9	14.1	17.7	22.6

R code:

```
> data=c(24.7,20.6,27.7,16.2,16.2,24.9,27.3,28.8,22.7,15.0,17.0,22.5,38.5,39.5,36.8,19.6,15.4,26.3,28.5,31.0,34.9,14.1,17.7,22.6)
> blocks=gl(4,6)
> treatments=gl(6,1,24)
> rbdfit=aov(data~blocks+treatments)
> rbdfit
Call:
   aov(formula = data ~ blocks + treatments)

Terms:
                blocks treatments Residuals
Sum of Squares  219.4279   901.1921  229.6396
Deg. of Freedom       3         5        15

Residual standard error: 3.912711
Estimated effects may be unbalanced
> summary.aov(rbdfit)
            Df Sum Sq Mean Sq F value   Pr(>F)
blocks       3  219.4   73.14   4.778   0.0157 *
treatments   5  901.2  180.24  11.773 9.28e-05 ***
Residuals   15  229.6   15.31
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interpretation :-

Here P< 0.05.Then Blocks are not homogenous .Treatments effects are not alike .

Latin Square Design (in Design of Experiments): The conducting of an experiment on r^2 experimental units, which differ in quality with respect to two characters and hence stratified, once with respect to one changing character, into different within-homogeneous rows and again, with respect to second changing character, into different within-homogeneous columns arranged in r^2-square form and then allotting 'r' treatments (whose effects are to be experimented) denoted by the Latin letters A, B, C, D, . . . to such within-homogenized units in each row separately without repetitions and also in each column separately without repetitions, splitting total variation in results into row-factor variation, column-factor variation, treatments variation, residual (error) variation and lastly testing the significance of these variations over error variation (variance).

In symbols, an experiment in which

Total variation =

Row-factor variation + Column-factor variation + Treatments variation+ Error variation

and then these variations are tested for their significance over error variation obtained from the results on applying r-treatments on r^2-units, without repetitions in any row or column, arranged in a square, which differ in their quality with respect to two characters, whose specimen with the treatments denoted in the Latin letters A, B, C & D, is given below:

A	B	C	D
B	C	D	A
C	D	A	B
D	A	B	C

Anova table for L.S.D:

Source of vari-ation	df	SS	Mean SS $=\left(\frac{SS}{df}\right)$	F-ratio (F-distribution)
Row-factor	r−1	$\sum_{i=1}^{r}\left(\frac{R_i^2}{r}\right)$ − CF	$\left(\frac{\text{Row SS}}{\text{r-1}}\right)$	$\left(\frac{\text{Row M SS}}{\text{Error MSS}}\right) \rightarrow F_{[r-1,\ (r-1)(r-2)\}}$
Column factor	r−1	$\sum_{j=1}^{r}\left(\frac{C_j^2}{r}\right)$ − CF	$\left(\frac{\text{Column SS}}{\text{r-1}}\right)$	$\left(\frac{\text{Column M SS}}{\text{Error MSS}}\right) \rightarrow F_{[r-1,\ (r-1)(r-2)\}}$
Treat-ments	r−1	$\sum_{k=1}^{r}\left(\frac{T_k^2}{r}\right)$ − CF	$\left(\frac{\text{Treatments SS}}{\text{r-1}}\right)$	$\left(\frac{\text{Treatments M SS}}{\text{Error MSS}}\right) \rightarrow F_{[r-1,\ (r-1)(r-2)\}}$
Error	(r−1)(r−2)	By subtraction	$\left(\frac{\text{Error SS}}{(r-1)(r-2)}\right)$	-----
Total	r^2-1	$\sum_{i=1}^{r}\sum_{j=1}^{r}\sum_{k=1}^{r} {x_{ijk}}^2$ − CF	-----	

In the above table, df = degrees of freedom ,SS = sum of squares, MSS = Mean SS = $\left(\frac{SS}{df}\right)$, CF = Correction factor = $\frac{(\text{Grand Total})^2}{\text{Total no. of observations}}$.

The F-ratio = $\left(\frac{\text{Row/Column/Treatment MSS}}{\text{Error MSS}}\right) \rightarrow F_{[r-1,\ (r-1)(r-2)\}}$ -distribution with (r–1) as the numerator df and (r–1)(r–2) as the denominator (Error) df. In the above ANOVA table, row SS +column SS + treatments SS + Error SS = total SS, but their MSS do not add together to give total MSS.

Latin Square Design:-

Suppose we want to analyze the productivity of 5 kinds of fertilizers, 5 kinds of tillage(land under cultivation) and 5 kinds of seeds. The data is organized in a LSD as follows

	Treat A	***Treat B***	***Treat C***	***Treat D***	***Treat E***
Fertilizer 1	***"A42"***	***"C47"***	***"B55"***	***"D51"***	***"E44"***
Fertilizer 2	***"E45"***	***"B54"***	***"C52"***	***"A44"***	***"D50"***
Fertilizer 3	***"C41"***	***"A46"***	***"D57"***	***"E47"***	***"B48"***
Fertilizer 4	***"B56"***	***"D52"***	***"E49"***	***"C50"***	***"A43"***
Fertilizer 5	***"D47"***	***"E49"***	***"A45"***	***"B54"***	***"C46"***

R code:

```
> fertil <- c(rep("fertil1",1), rep("fertil2",1), rep("fertil3",1), rep("fertil4",1), rep("fertil5",1))
> treat <- c(rep("treatA",5), rep("treatB",5), rep("treatC",5), rep("treatD",5), rep("treatE",5))
> seed <- c("A","E","C","B","D", "C","B","A","D","E", "B","C","D","E","A", "D","A","E","C","B", "E","D","B","A","C")
> 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)
> mydata <- data.frame(treat, fertil, seed, freq)
> myfit <- lm(freq ~ fertil+treat+seed, mydata)
> anova(myfit)
```

```
Analysis of Variance Table

Response: freq
          Df Sum Sq Mean Sq F value    Pr(>F)
fertil     4  17.76   4.440  0.7967 0.549839
treat      4 109.36  27.340  4.9055 0.014105 *
seed       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
```

Interpretation:

– The difference between group considering the fertilizer is not significant (p-value > 0.1);
– The difference between group considering the tillage is quite significant (p-value < 0.05);
– The difference between group considering the seed is very significant (p-value < 0.001);

Practice Questions and Challenging Experiments:-

Completely randomised design:-

1. A firm wishes to compare four programs for training workers to perform a certain manual task. Twenty new employees are randomly assigned to the training programs, with 5 in each program. At the end of the training period, a test is conducted to see how quickly trainees can perform the task. The number of times the task is performed per minute is recorded for each trainee

Program 1	Program 2	Program 3	Program 4
9	10	12	9
12	6	14	8
14	9	11	11
11	9	13	7
13	10	11	8

Calculate and interpret the above one way ANOVA table.

2. In a factory producing edible oil and marketing its product in 15 kg tins, uses five filling machines. Random samples of the packed tins were taken for each machine A,B,C,D and E were presented below

A	B	C	D	E
14.85	14.28	14.16	15.25	14.60
15.00	14.42	14.15	15.30	14.84
15.25		14.19	15.10	14.82
15.10		14.50	15.35	14.74
14.80			15.00	

Analysis of data to test the Equality of efficiency of machines .

Randomised Block Design (R.B.D)

A factory manager wanted to compare the efficiency of four factory workers with respect to cotton spinning. He had four machines .Four workers A,B,C and D were allotted a machine as experimental units. Cotton thread yield(in kg) for each worker

was recorded as shown in the layout displayed below. All machines were of the same make

Blocks	Machines				
I	C(22)	A(14)	B(12)	A(23)	
II	A(16)	B(18)	C(20)	D(25)	
III	A(15)	C(23)	D(28)	B(14)	
IV	A(17)	D(21)	C(19)	B(11)	
V	B(18)	C(20)	A(13)	D(24)	
VI	D(18)	C(21)	A(10)	B(17)	

Latin Square Design (L.S.D)

Analyse and interpret the following statistics concerning output of wheat per field obtained as a result of experiment conducted to test four varieties of wheat Viz., A,B,C and D under a Latin Square Design

C (25)	**B (23)**	**A(20)**	**D (20)**
A (19)	**D (19)**	**C (21)**	**B (18)**
B (19)	**A (14)**	**D (17)**	**C (20)**
D (17)	**C (20)**	**B (21)**	**A (15)**