Chapter 19: R Programming

# Learning Objectives

After reading this chapter, you will be able to

* Understand the basic data types used in R
* Analyse matrix operations
* Understand the purpose of list
* Know the syntax of data fame
* Implement data import
* Construct frequency distribution of quantitative data using COUNT function
* Analyse data using histogram
* Construct cumulative frequency distribution of ungrouped data and grouped data
* Construct cumulative frequency graph
* Analyse numerical measures, namely mean median, quartile, percentile, range, interquartile range, variance, standard deviation, covariance, correlation coefficient, skewness and kurtosis
* Understand probability distributions
* Determine interval estimate
* Analyse hypotheses testing
* Conduct goodness-of-fit test and analyse categorical data using chi-square test
* Analyse data using different designs of ANOVA
* Implement different non-parametric tests
* Analyse data using simple and multiple regression models

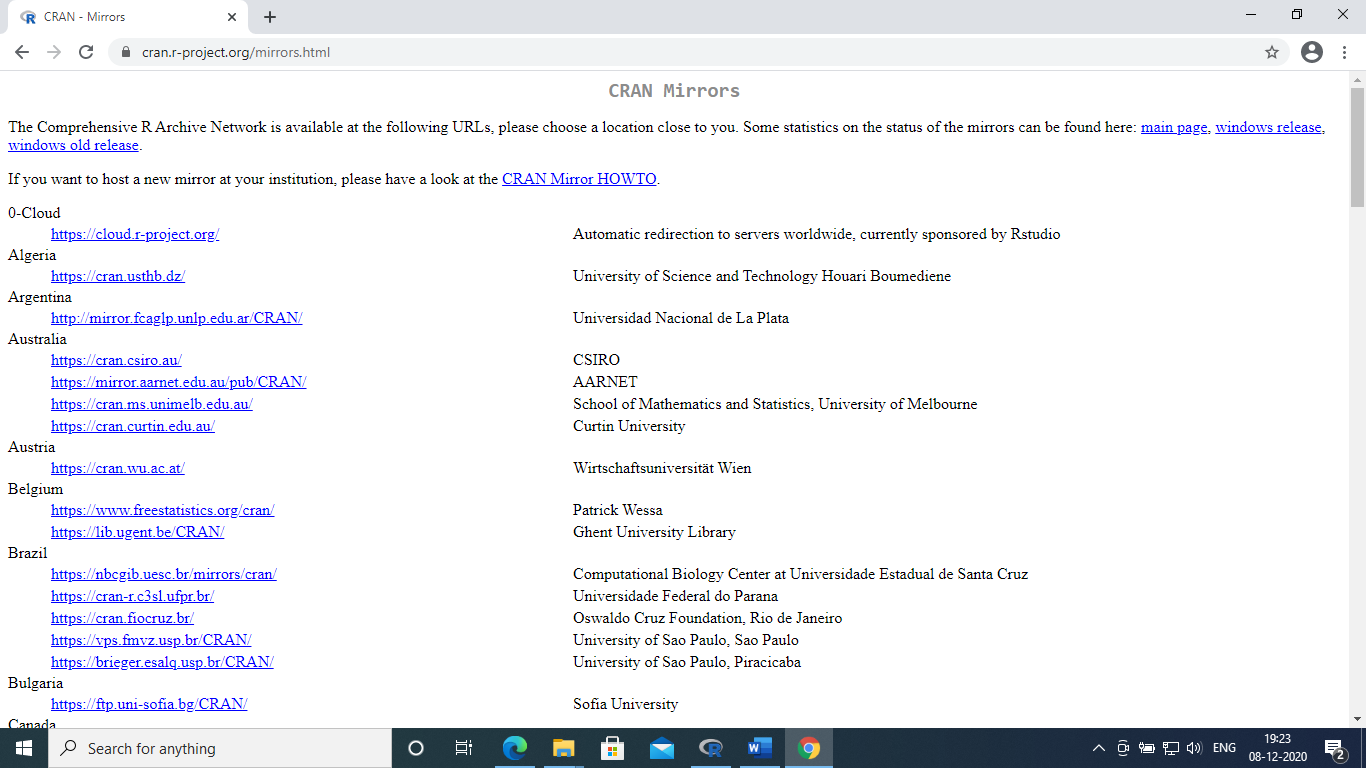
# 19.1. Introduction

R programming is a programming software, which has the feature of having functions for most of the modules of computing, namely vector, matrix multiplication, measure of central tendencies, regression model and so on. When this is compared to any high-level language programming for statistics, the extent of code to be written is very minimal. Further, R software is an open-source software, which can be used by any researcher/practitioner.

**Installation of R Software:**The sequence of steps involved in installing R software is presented below.

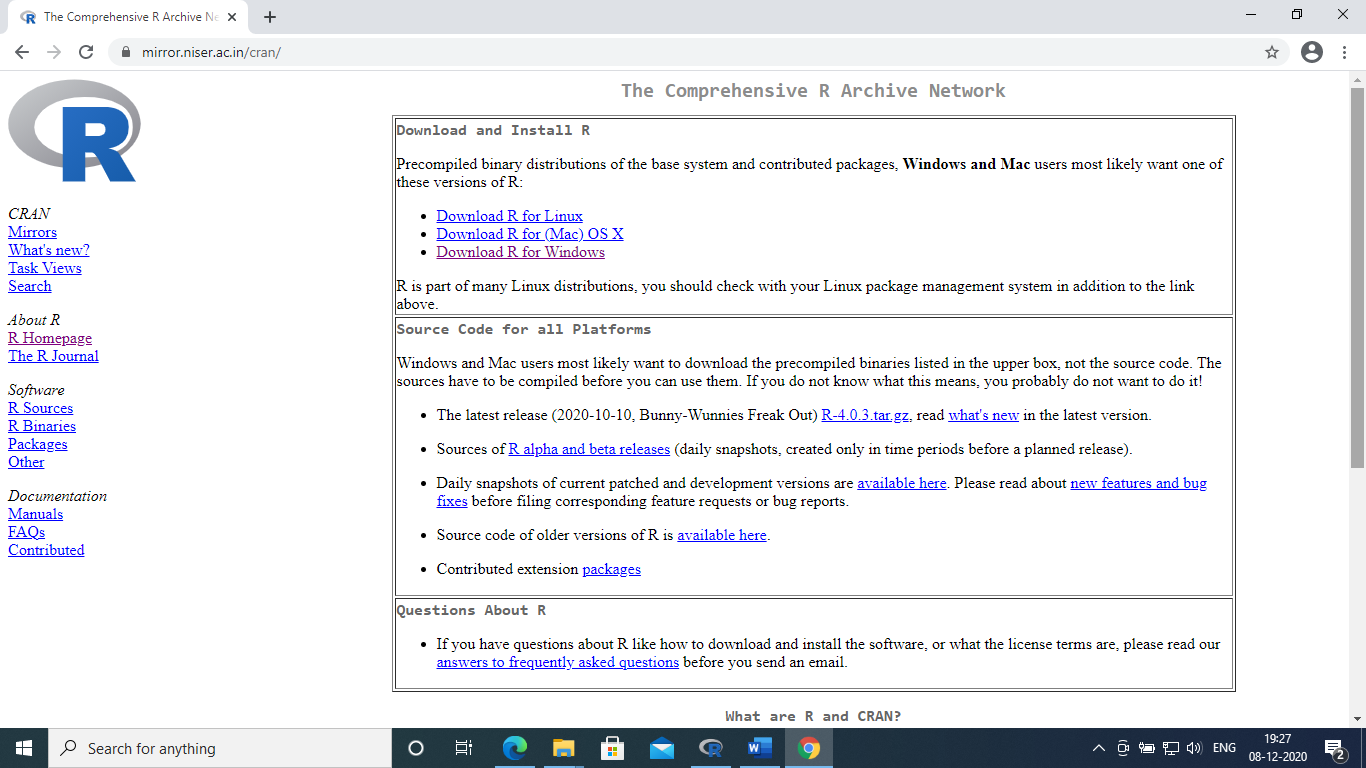
**Step 1:** Enter cran.r-project.org › mirror in search menu bar of google and click address.

**Step 2:** The response screen for the action in Step 1 is shown in Figure 19.1.



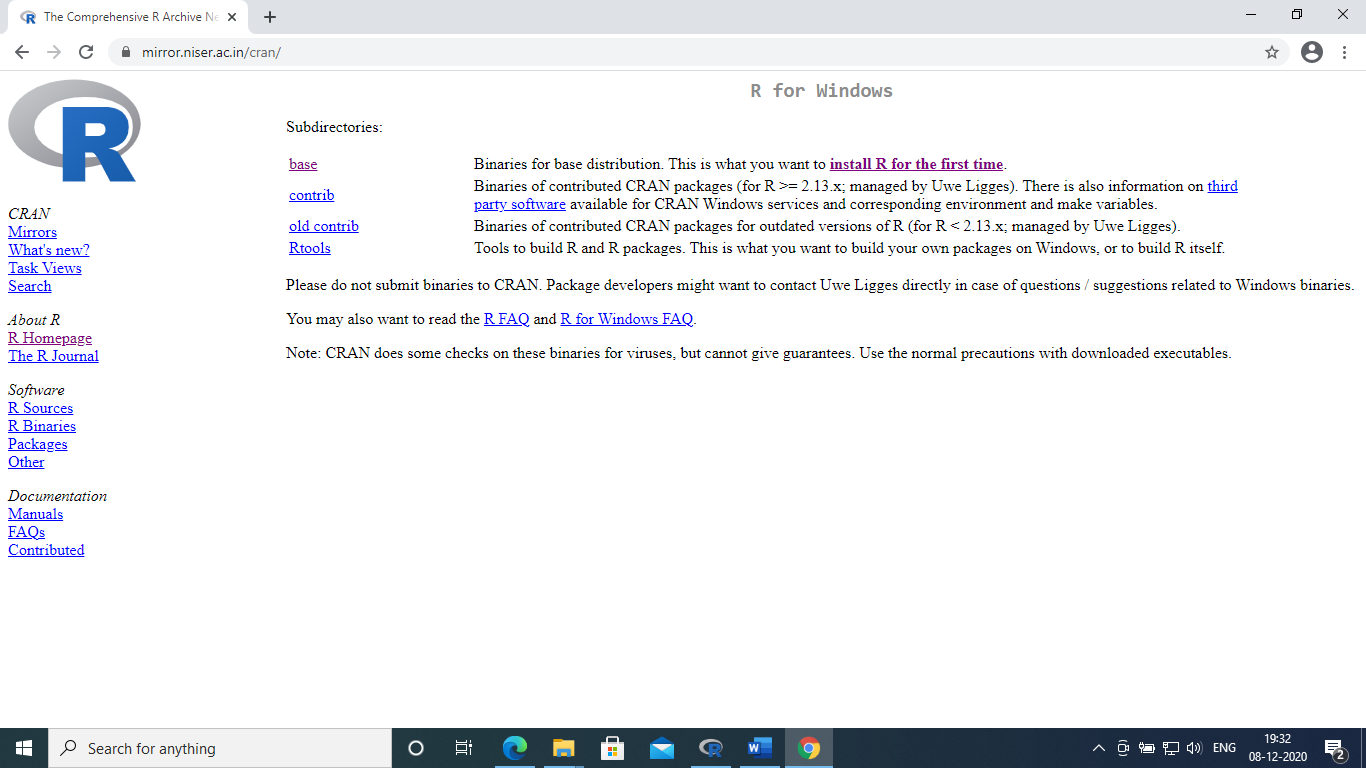
**Figure 19.1.** Screenshot for the Action in Step 1

**Step 3:** Scroll down the cursor to locate your nearest location (country) and click that location, which gives a display as shown in Figure 19.2.



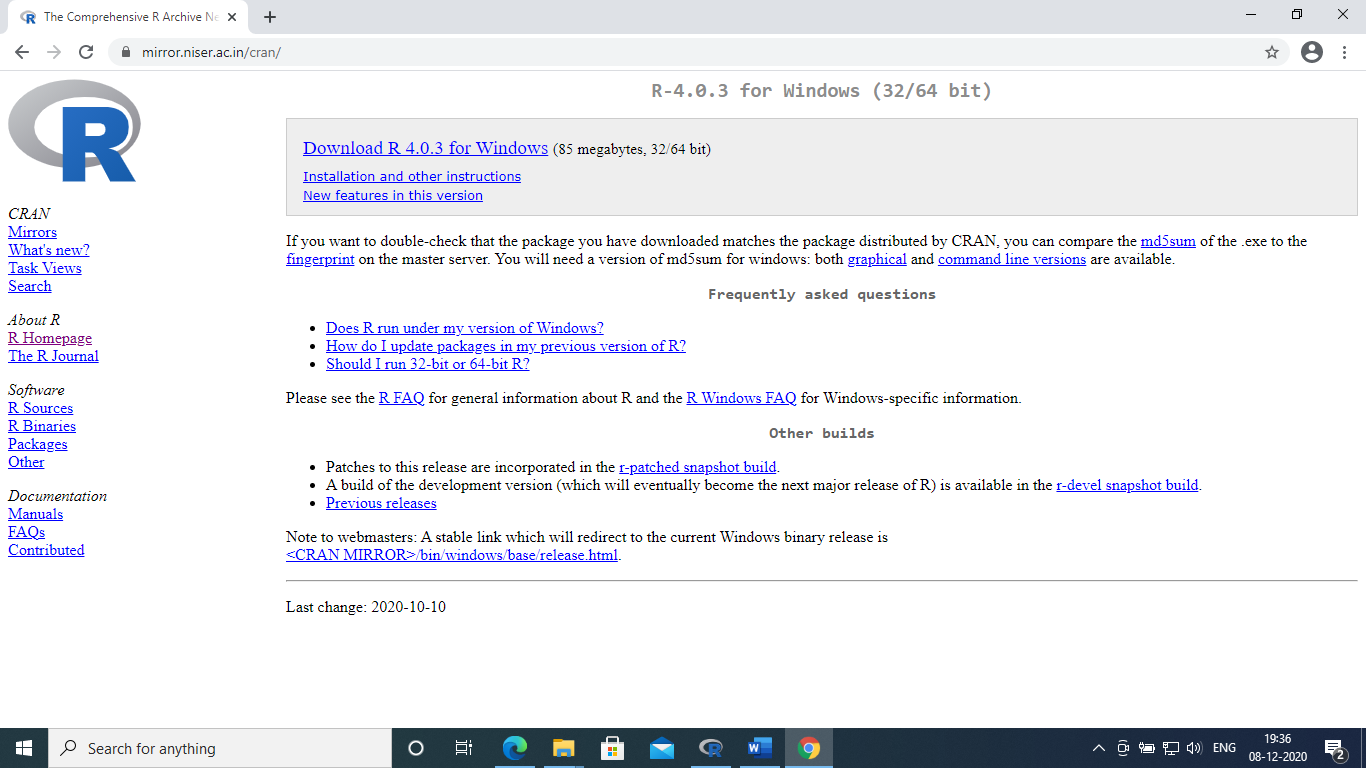
**Figure 19.2.** Screenshot after Clicking the Country

**Step 4:** Click the option "Download R for Window", which gives a display as shown in Figure 19.3.



**Figure 19.3.** Screenshot after Clicking the Operating System in the Menus of Figure 19.2

**Step 5:** Click "Install R for the First Time" in the first line in the screenshot of Figure 19.3, which gives a display as in Figure 19.4.



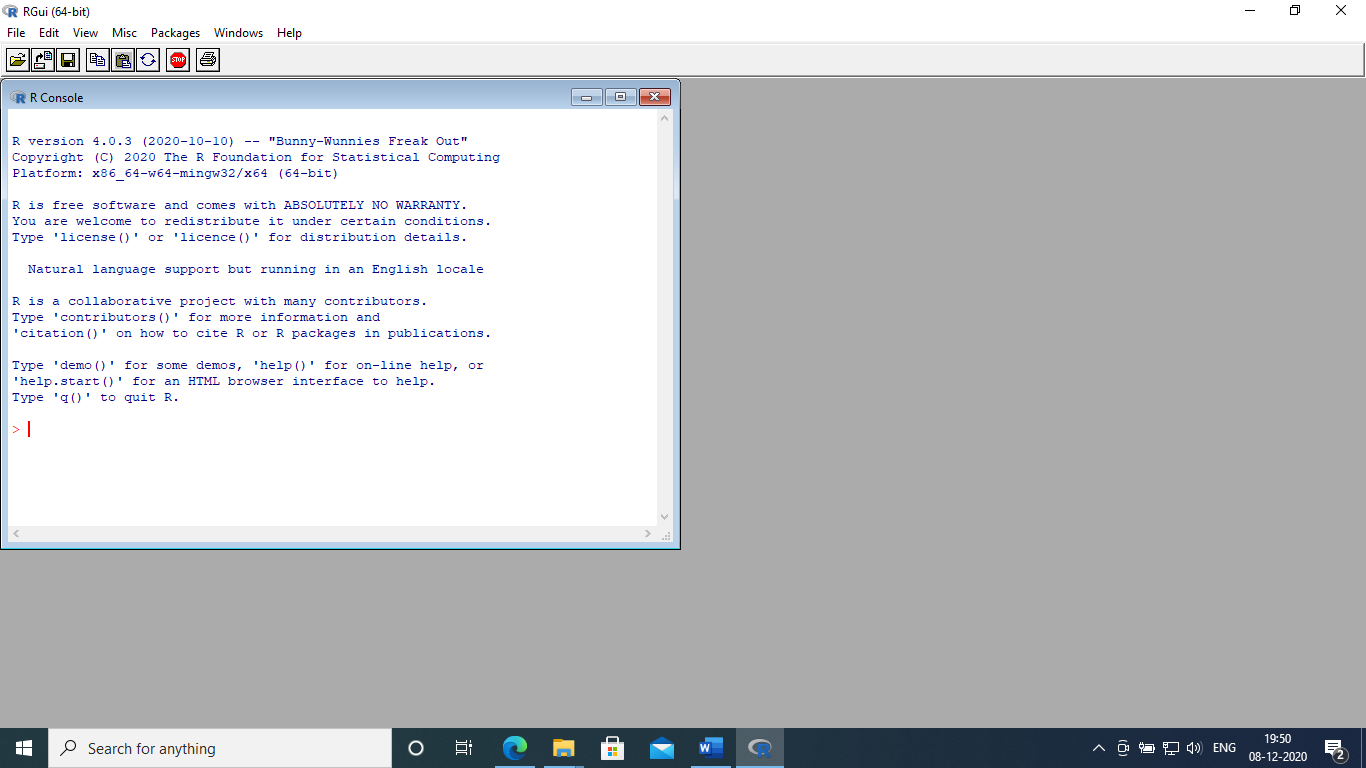
**Figure 19.4.** Screenshot after Clicking "Install R for the First Time" in Figure 19.3

**Step 6:** Click "Download R.4.0.3 for Windows" in the screenshot shown in Figure 19.4, which downloads a file "R-4.0.3-win.exe".

**Step 7:** Click the downloaded file to install R software in your computer with 32 bits or 64 bits.

# 19.2. Introduction to R

The click of desktop icon of R software gives a display as in Figure 19.5. The statements of R programming are to be entered after the ">" prompt shown in Figure 19.5.



**Figure 19.5.** Screenshot after Clicking Desktop Icon of R Software

## 19.2.1. Use of External Data in R

In R programming, the data can be fed through keyboard as explained below.

### 19.2.1.1. Adding Two Numbers at Prompt Level

At the prompt level, enter the following to add two numbers, which gives the result in the next line of the cursor.

> 10 + 20 # Adding two numbers

[1] 30

>

### 19.2.1.2. Assignment Statement

The external data that are to be given as input are assigned to variables, which can be used for number crunching operations later according to programming requirements.

> weight = 10 # Assignment statement

> weight # Printing weight

[1] 10

>

*Adding two numbers stored in variables*

> *x* = 15

> *y* = 20

> *x* + *y* # Adding two number

[1] 35

>

### 19.2.1.3. Function

In R, function is used to achieve a task. If four numbers are to be treated as a vector, then a combine function c is used as shown below. The function has a name, open parenthesis and zero or more arguments and close parenthesis.

> c(5, 2, 3, 8) # Defining vector

[1] 5 2 3 8

>

### 19.2.1.4. Square Root of a Number

The square root of a number is obtained using sqrt function as shown below.

> sqrt(100)

[1] 10

>

### 19.2.1.5. Getting Help for Partly Known Commands

If the function or command of interest is not known fully, then one can use fuzzy search using apropos command as shown below.

> apropos("ano") # Help command

[1] "anova" "mahalanobis" "manova" "power.anova.test"

[5] "stat.anova" "summary.manova" "volcano"

>

# 19.3. Basic Data Types

The data types handled in R programming are listed below.

1. Numeric
2. Integer
3. Complex
4. Logic
5. Character

## 19.3.1. Handling Numeric Data

The data with decimal digits or without decimal digits are called numeric data in R. The usage of a numeric data is shown below.

*Numeric data with decimal digits*

> *x* = 20.768 # Defining numeric data

> *x*

[1] 20.768

*Numeric data without decimal digits*

In the following case, z is stored with 5, which is a numeric data. The class to which it belongs can be identified using *class* command as shown below the first part of displacing z.

# Part 1 Comment statement

> z = 5 # Defining numeric data

> z

[1] 5

# Part 2

> class(z) # Checking class of z

[1] "numeric"

>

## 19.3.2. Handling Integer Data

The syntax with data of defining a number as an integer is shown below. The class to which the integer variable *p* belongs is shown immediately after the first part.

>#Part 1 Comment statement

> z = 5

> class(z) # Checking class

[1] "numeric"

> p = as.integer(234) # Defining integer

># Part 2

> class(p) # Checking class of p

[1] "integer"

>

Further, one can check whether variable *p* is integer type or not as shown below, which gives either TRUE if it is integer or FALSE otherwise.

> is.integer(p) # Checking whether p is integer

[1] TRUE

>

### 19.3.2.1. Truncation of Decimal Number into Integer Number

A decimal number is truncated into integer number after dropping the decimal digits as shown below using as.integer function.

> as.integer(25.346) # Printing integer of a number

[1] 25

>

**19.3.2.2. Value for TRUE or FALSE**

The value for TRUE or FALSE is obtained as shown below in R.

> as.integer(TRUE) # Value of TRUE

[1] 1

> as.integer(FALSE) # Value of FALSE

[1] 0

>

## 19.3.3. Complex Data

Complex data can be represented in R using real and imaginary components of the complex number as shown below. The class to which the complex variable *x* belongs is checked in the second part of the following program.

#Part 1

> x = 10 + 3i # Defining complex variable

> x

[1] 10 + 3i

#Part 2

> class(x) # Printing type of class of variable

[1] "complex"

**>**

## 19.3.4. Logical Data

When two values are compared, the result may be either true or false, which belongs to logical data.

Example: If a > b then TRUE; else FALSE. In each statement, text after # forms comment of the statement.

> p = 12; q = 8 # assignment of values to variables

> x = p > q # comparison of p and q and storing logical value in x

> x # printing logical value

[1] TRUE

>

The other logical operators are listed below.

The logical operator AND is given by &.

The logical operator OR is given by |.

The logical operator Negation (NOT) is given by .

These are explained below.

> x = TRUE; y = FALSE

> x & y # AND operation of logical variables

[1] FALSE

> x|y # OR operation of logical variables

[1] TRUE

> !x # NOT (negation) operation of logical variables

[1] FALSE

>

## 19.3.5. Character Data

Character function is used to convert an object into string. The function for character is as.character().

A sample treatment of an object as a character in R is given below. The second part of the program gives the status of variable *p*.

#Part 1

> p = as.character(20.3455) # Printing as character data

> p

#Part 2

[1] "20.3455"

> class(p)

[1] "character"

>

### 19.3.5.1. Concatenation of Characters

The concatenation of strings is performed using paste function in R, which is demonstrated using the following R program.

> m = "John"; n = "Arrived"

> paste(m, n) # merging two strings using paste function

[1] "John Arrived"

>

### 19.3.5.2. Extracting Substring from a String

A substring from a string can be extracted using substr function.

> substr("Dividend Rate is 20 percent.", start = 10, stop = 27)

[1] "Rate is 20 percent."

>

# 19.4. Vector

A vector consists of a sequence of elements of data of the same type such as numeric and integer. The function to represent a vector is c(). A sample definition of a vector is shown below in R.

*Vector with integer data type*

> c(10, 8, 12, 25) # Printing vector with integer data

[1] 10 8 12 25

>

*Vector with character data type*

> c("Square", "Rectangle", "Square") # Printing vector with character data

[1] "Square" "Rectangle" "Square"

>

*Vector with logical data*

> c(TRUE, FALSE, TRUE, TRUE) # Printing vector with logical data

[1] TRUE FALSE TRUE TRUE

>

## 19.4.1. Length of Vector

The length of a vector can be obtained using the syntax length(c()), which represents the number of members in the vector.

> length(c(12, 8, 19, 34, 22)) # Printing length of vector

[1] 5

>

## 19.4.2. Vector Arithmetic

The addition, subtraction, multiplication and division of the vectors can be carried out using the respective arithmetic operators. These are demonstrated in the following R program using two vectors x and y.

> x = c(5, 3, 6, 10); y = c(2, 1, 3, 5)

> x + y # Addition of two vectors of equal size

[1] 7 4 9 15

> x – y # Subtraction of two vectors of equal size

[1] 3 2 3 5

> x × y # Multiplication of two vectors of equal size

[1] 10 3 18 50

> x/y # Division of two vectors of equal size

[1] 2.5 3.0 2.0 2.0

>

## 19.4.3. Recycling Rule Applied to Vector Operations

The arithmetic operations (+, –, ×, /) can be applied to two vectors. While doing so, the two vectors should have equal size, which is achieved using recycling rule. If vector p = (2, 4, 1, 3, 6, 5) and q = (4, 3), then equal size for vector y is obtained by repeating the members of y till the size of vector y is made equal to vector x. One should note the fact that the vector with small size is recycled. The process of recycling is automatic, when a particular arithmetic operation is carried out on two vectors. The different arithmetic operations applied to the two vectors, namely p and q using the recycling operation in R program, are shown below.

> p = c(2, 4, 1, 3, 6, 5)

> q = c(4, 3)

> p + q # Addition of two vectors with cycling

[1] 6 7 5 6 10 8

> p – q # Subtraction of two vectors with cycling

[1] –2 1 –3 0 2 2

> p × q # Multiplication of two vector with cycling

[1] 8 12 4 9 24 15

> p/q # Division of two vectors with cycling

[1] 0.500000 1.333333 0.250000 1.000000 1.500000 1.666667

>

## 19.4.4. Vector Index

In R program, vector index retrieves an element from the vector using its position in the vector. The vector index is represented using square brackets [ ] with the position of the element in them, which is to be retrieved. In a way, it is equivalent to a subscripted variable in high-level languages.

> a = c("pp", "qq", "rr", "xx", "yy", "zz") # Defining vector *a*

> a[4] # printing a member of vector using vector index

[1] "xx"

>

### 19.4.4.1. Negative Indexing

Negative indexing removes the member of a vector corresponding to an index as shown in the following R program. The index "–3" removes the third element from vector "a" and gives the remaining elements in their order.

> a = c("pp", "qq", "rr", "xx", "yy", "zz") # Defining negative indexing

> a[–3]

[1] "pp" "qq" "xx" "yy" "zz"

>

## 19.4.5. Numeric Index Vector

The numeric index vector fetches a subset of the given vector as shown below.

> m = c("aa", "bb", "cc", "dd", "ee", "ff")

> m[c(2, 4, 6)] # Defining numeric indexing vector

[1] "bb" "dd" "ff"

>

In this program, the string in locations 2, 4 and 6 are retrieved from vector *m* using numeric index vector.

### 19.4.5.1. Duplicates Indexing

If the same position of a vector is repeated in a numeric indexing vector, that particular element will be included once again. A sample R program is shown below.

> m = c("aa", "bb", "cc", "dd", "ee", "ff") # Vector m

> m[c(2, 4, 4)] # Printing vector members with duplications

[1] "bb" "dd" "dd"

>

### 19.4.5.2. Range Indexing

In range indexing, the range of location numbers of a vector will retrieve the members of the vector in that range. A sample R program is shown below.

> p = c(10, 20, 15, 28, 19, 45) # Defining vector c

> p[2:4] # Defining range indexing

[1] 20 15 28

>

## 19.4.6. Logical Index Vector

The elements of a vector can be retrieved by including TRUE in the corresponding location of logical index vector. If an element in a vector is not to be retrieved, then FALSE is to be included in the corresponding location of that element. The following R program demonstrates this concept.

> p = c(1, 2, 3, 4, 5, 6)

> L = c(TRUE, TRUE, FALSE, FALSE, TRUE, FALSE) # Vector of logical variables

> p[L]

[1] 1 2 5

In the above program, line 2 and line 3 can be combined as shown below.

> p = c(1, 2, 3, 4, 5, 6) # Printing members of vector using logical variables

> p[c(TRUE, TRUE, FALSE, FALSE, TRUE, FALSE)]

[1] 1 2 5

>

## 19.4.7. Named Vector Members

The members of a vector can be named, which will help to give some identification for that member.

> p = c("Peter", "Queen", "Varun") # Defining and printing members of vector c

> p

[1] "Peter" "Queen" "Varun"

> names(p) = c("First Name", "Second Name", "Third Name")

> p # Printing the members of vector with names

First Name Second Name Third Name

"Peter" "Queen" "Varun"

>

In the above R program, vector p is defined and displayed first. Then the members, namely Peter, Queen and Varun in vector p, are named as First Name, Second Name and Third Name, respectively.

Specifically, if an element is to be retrieved using its name, then that name is to be used as the index of that vector as shown in the following R programming based on the vector, which has been already define above.

> p["Second Name"] # Printing vector element "Queen" with its name

Second Name

"Queen"

>

Similarly, the elements of vector "c" can be retrieved in any order using their names as shown below.

> p[c("Third Name", "Second Name", "First Name")] # Printing vector c with names for members

Third Name Second Name First Name

"Varun" "Queen" "Peter"

>

# 19.5. Matrix

A matrix is a two-dimensional array having a specified number of rows and a specified number of columns. If the number of rows is 2 and the number of columns is 4, then the total number of data elements stored in the matrix is 8.

The matrix function in R program is matrix (). In R program to define a matrix p, a vector c, number of rows, number of columns and the order of presenting the data, namely byrow = TRUE or bycol = TRUE are included within the parentheses of matrix.

These can be written with comments as given below.

> p = matrix(

+ c(1, 2, 3, 4, 5, 6, 7, 8), # Matrix elements in vector

+ nrow = 2, # No. of rows

+ ncol = 4, # No. of columns

+ byrow = TRUE) # matrix filling by rows

> p # printing matrix p

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

[1,] 1 2 3 4

[2,] 5 6 7 8

>

The above program can be rewritten by omitting comments as shown below.

> p = matrix(c(1, 2, 3, 4, 5, 6, 7, 8), nrow = 2, ncol = 4, byrow = TRUE) # Defining a matrix

> p

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

[1,] 1 2 3 4

[2,] 5 6 7 8

>

Element 7 at the second row and third column of matrix p can be referenced using its row number and column number as shown below.

> p[2,3] # Printing a specific matrix element

[1] 7

>

All the elements in a row can be displayed using "row\_number,". The elements in row 2 of matrix p are shown in the following R program.

> p[2,] # Printing all the elements of row 2 of matrix p

[1] 5 6 7 8

>

All the elements in a column can be displayed using ",column\_number,". The elements in column 3 of matrix p are shown in the following R program.

> p[,3] # Printing all the elements of column 3 of matrix p

[1] 3 7

>

The following R program prints rows 1 and 2 of matrix p.

> p[c(1, 2),] # Printing all the elements of rows 1 and 2

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

[1,] 1 2 3 4

[2,] 5 6 7 8

>

The following R program displays columns 2 and 4 of matrix p.

> p[ ,c(2, 4)] # Printing all the elements of columns 2 and 4

[,1] [,2]

[1,] 2 4

[2,] 6 8

>

The rows and columns of matrix p can be named as R1, R2, C1, C2, C3 and C4, respectively, using dimnames functions as shown below.

> dimnames(p) = list(

+ c("R1", "R2"), #Row names

+ c("C1", "C2", "C3", "C4")) # Column names

> p

C1 C2 C3 C4

R1 1 2 3 4

R2 5 6 7 8

>

The same program is written with row and column names as shown below.

> dimnames(p) = list(c("R1", "R2"), c("C1", "C2", "C3", "C4")) # Naming rows and columns  
of matrix

> p # printing matrix p

C1 C2 C3 C4

R1 1 2 3 4

R2 5 6 7 8

>

A matrix element can be accessed using the names and columns of that element as shown below.

> p["R1", "C4"] # Printing a matrix element

[1] 4

>

# 19.6. List

List contains more than one vector. These vectors can be referenced using indexing concept. In the following R program, three vectors, namely *a*, *b* and *d*, are included in the list called *p*. Then, vector *a* and vector *d* are displayed using index of list *p* using p[c(1, 3)], because vector *a* and vector *d* are the first and third vectors, respectively, in list *p*.

> a = c(1, 2, 4, 5)

> b = c("Accept", "Reject")

> d = c("Good", "Very Good", "Excellent")

> p = list(a, b, d) # Defining list

> p[c(1, 3)] # Printing member 1 and member 3 of list

[[1]]

[1] 1 2 4 5

[[2]]

[1] "Good" "Very Good" "Excellent"

## 19.6.1. Member Reference

The member of vector included in a list can be accessed using double indexing. The second member in the third vector included in list *p* is obtained as shown in the following R program.

> p[[3]][2] #Referencing third row and second column of p

[1] "Very Good"

>

Accessing the entire vector

The entire vector can be accessed using double square brackets as shown below to access the second vector.

> p[[2]] # Accessing members of entire list

[1] "Accept" "Reject"

>

## 19.6.2. Modification of the Content of a Member Temporally

The content of a member of a vector included in a list can be modified temporally as shown in the following R program.

> p[[3]][1] ="Moderate" # Temporary modification of member of list

> p[[3]]

[1] "Moderate" "Very Good" "Excellent"

> d # Printing content of original vector

[1] "Good" "Very Good" "Excellent"

>

In the first part of the above R program, the first member of the third vector in list p is modified to "Moderate", which is displayed along with other members of that vector using the command p[[3]].

The second of the R program shows that the contents in vector d, which is the third vector of list p, are unaltered.

# 19.7. Data Frame

In R, data frame is used to store data table. The data frame consists of a list of vectors of equal length. Consider the problem of storing sales data of salesmen in four different quarters as shown in Table 19.1.

**Table 19.1.** Quarterly Sales Data (Crores of Rupees)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Salesman |  | Period | | | |
|  | Quarter 1 | Quarter 2 | Quarter 3 | Quarter 4 |
| 1 | 20 | 22 | 15 | 18 |
| 2 | 22 | 18 | 24 | 15 |
| 3 | 30 | 25 | 17 | 23 |
| 4 | 25 | 28 | 20 | 23 |

The sales of each salesman are treated as a vector with four members (quarterly sales).

The definition of the vectors using data frame command data.frame( ) and manipulation of the data are shown in the following R program.

>sm1 = c(20, 22, 15, 18)

> sm2 = c(22, 18, 24, 15)

> sm3 = c(30, 25, 17, 23)

> sm4 = c(25, 28, 20, 23)

> s = data.frame(sm1, sm2, sm3, sm4) # Defining dataframe

> s # Printing content of datafame

sm1 sm2 sm3 sm4

1 20 22 30 25

2 22 18 25 28

3 15 24 17 20

4 18 15 23 23

In the above program, modifications are incorporated to have names of rows and columns as shown below. In the data frame, the salesmen are kept in columns and the quarters are kept in rows.

> dimnames(s) = data.frame(c("Qtr1","Qtr2","Qtr3","Qtr4"), c("SM1","SM2","SM3","SM4"))

+ # Naming rows and columns of dataframe

> s # Printing content of dataframe

SM1 SM2 SM3 SM4

Qtr1 20 22 30 25

Qtr2 22 18 25 28

Qtr3 15 24 17 20

Qtr4 18 15 23 23

>

The number of rows and the number columns of the data frame are displayed in the following program.

> ncol(s) # Printing number of columns of dataframe

[1] 4

> nrow(s) # Printing number of rows of dataframe

[1] 4

> ncol(s) # Printing number of columns of dataframe

[1] 4

>

A particular data element in the data frame is displaced as shown below in R program. The value of the data element in Quarter 2 for salesman 1 is displayed in the following R program.

> s["Qtr2", "SM1"]

[1] 22

>

## 19.7.1. Data Frame Column Vector

The data items in a data frame can be accessed in different ways.

The data frame *s* of the data shown in Table 19.1 can be accessed in different ways as shown below.

> sm1 = c(20, 22, 15, 18)

> sm2 = c(22, 18, 24, 15)

> sm3 = c(30, 25, 17, 23)

> sm4 = c(25, 28, 20, 23)

> s = data.frame(sm1, sm2, sm3, sm3)

> dimnames(s) = data.frame(c("Qtr1", "Qtr2", "Qtr3", "Qtr4"), c("SM1", "SM2", "SM3", "SM4"))

> s[["SM1"]] # using square brackets

[1] 20 22 15 18

> s$SM1 # Using $ symbol

[1] 20 22 15 18

> s[,"SM1"] # Using ,"SM1"

[1] 20 22 15 18

>

In the above R program, three types of accessing the members of the vector of salesman 1, SM1, are included, which are presented below.

> s[["SM1"]] # Using Square brackets

> s$SM1 # Using $ symbol

> s[,"SM1"] # Using ,"SM1"

## 19.7.2. Data Frame Column Slicing

The data items in a data frame can be accessed in terms of scaling the vectors in it in different ways.

The vector of the data frame *s* of the data shown in Table 19.1 can be sliced in different ways as shown below.

> sm1 = c(20, 22, 15, 18)

> sm2 = c(22, 18, 24, 15)

> sm3 = c(30, 25, 17, 23)

> sm4 = c(25, 28, 20, 23)

> s = data.frame(sm1, sm2, sm3, sm3)

> dimnames(s) = data.frame(c("Qtr1", "Qtr2", "Qtr3", "Qtr4"), c("SM1", "SM2", "SM3", "SM4"))

### 19.7.2.1. Slicing of Data Frame Using Numeric Indexing

The following R programming statement gives the members of vector sm1.

> s[1]

SM1

Qtr1 20

Qtr2 22

Qtr3 15

Qtr4 18

>

### 19.7.2.2. Slicing of Data Frame Using Name Indexing

The following R program statement gives the members of vector sm1 using its name "SM1".

> s["SM1"]

SM1

Qtr1 20

Qtr2 22

Qtr3 15

Qtr4 18

### 19.7.2.3. Slicing Multiple Columns

The following R program statement gives the members of two vectors "SM1" and "SM2".

> s[c("SM1", "SM2")]

SM1 SM2

Qtr1 20 22

Qtr2 22 18

Qtr3 15 24

Qtr4 18 15

>

## 19.7.3. Data Frame Row Slicing

A select row can be sliced from a data frame in R. Considering the vectors sm1, sm2, sm3 and sm4, which have been already defined in the data frame *s* along with names for the columns and rows of the data frame presented earlier, an R program to slice row 1, which corresponds to Qtr1, is presented using numeric indexing.

>sm1 = c(20, 22, 15, 18)

> sm2 = c(22, 18, 24, 15)

> sm3 = c(30, 25, 17, 23)

> sm4 = c(25, 28, 20, 23)

> s = data.frame(sm1, sm2, sm3, sm3)

> dimnames(s) = data.frame(c("Qtr1", "Qtr2", "Qtr3", "Qtr4"), c("SM1", "SM2", "SM3", "SM4"))

### 19.7.3.1. Numeric Indexing

An R program to slice row 1, which corresponds to Qtr1, is presented using numeric indexing as shown in the following R program.

>sm1 = c(20, 22, 15, 18)

> sm2 = c(22, 18, 24, 15)

> sm3 = c(30, 25, 17, 23)

> sm4 = c(25, 28, 20, 23)

> s = data.frame(sm1, sm2, sm3, sm3)

> dimnames(s) = data.frame(c("Qtr1", "Qtr2", "Qtr3", "Qtr4"), c("SM1", "SM2", "SM3", "SM4"))

> s[1,]

SM1 SM2 SM3 SM4

Qtr1 20 22 30 30

>

An R program to slice row 1 and row 3, which correspond to Qtr1 and Qtr 3, respectively, using numeric indexing is shown below.

>sm1 = c(20, 22, 15, 18)

> sm2 = c(22, 18, 24, 15)

> sm3 = c(30, 25, 17, 23)

> sm4 = c(25, 28, 20, 23)

> s = data.frame(sm1, sm2, sm3, sm3)

> dimnames(s) = data.frame(c("Qtr1", "Qtr2", "Qtr3", "Qtr4"), c("SM1", "SM2", "SM3", "SM4"))

> s[c(1,3),]

SM1 SM2 SM3 SM4

Qtr1 20 22 30 30

Qtr3 15 24 17 17

>

### 19.7.3.2. Name Indexing

An R program for slicing of row 3, which corresponds to Qtr3 using name indexing, is shown below.

>sm1 = c(20, 22, 15, 18)

> sm2 = c(22, 18, 24, 15)

> sm3 = c(30, 25, 17, 23)

> sm4 = c(25, 28, 20, 23)

> s = data.frame(sm1, sm2, sm3, sm3)

> dimnames(s) = data.frame(c("Qtr1", "Qtr2", "Qtr3", "Qtr4"), c("SM1", "SM2", "SM3", "SM4"))

> s["Qtr3",]

SM1 SM2 SM3 SM4

Qtr3 15 24 17 17

>

An R program to slice row 1 and row 3, which correspond to Qtr1 and Qtr3, respectively, using name indexing is shown below.

>sm1 = c(20, 22, 15, 18)

> sm2 = c(22, 18, 24, 15)

> sm3 = c(30, 25, 17, 23)

> sm4 = c(25, 28, 20, 23)

> s = data.frame(sm1, sm2, sm3, sm3)

> dimnames(s) = data.frame(c("Qtr1", "Qtr2", "Qtr3", "Qtr4"), c("SM1", "SM2", "SM3", "SM4"))

> s[c("Qtr1", "Qtr3"),]

SM1 SM2 SM3 SM4

Qtr1 20 22 30 30

Qtr3 15 24 17 17

### 19.7.3.3. Logical Indexing

A logical indexing gives TRUE and FALSE instances corresponding to the members of a vector in a data frame based on a logical expression. First, a comparative statement, which gives TRUE or FALSE, is constructed. Then typing the left-hand side of the statement at prompt level and them pressing ENTER key gives the outcomes of the logical indexing of the members of the vector(s) in the data frame.

An R program to get TRUE or FALSE outcome for the members of the vector with name SM1 for the condition s$SM1 < 20 is shown below.

> L = s$SM1 < 20

> L

[1] FALSE FALSE TRUE TRUE

>

Based on the data frame shown below, an R program to get the rows of the data frame for the condition s$SM1 < 20 is presented below. In row 1 as well as in row 2, the value in column SM1 is greater than 20. Hence, row 1 (Qtr1) and row 2 (Qtr2) are not printed in the output. Row 3 (Qtr3) and row 4 (Qtr4) are printed, because the value in each of those rows in column SM1 is less than 20.

*Data Frame s*

SM1 SM2 SM3 SM4

Qtr1 20 22 30 25

Qtr2 22 18 25 28

Qtr3 15 24 17 20

Qtr4 18 15 23 23

*R Program*

> L = s$SM1 < 20

> s[L,]

SM1 SM2 SM3 SM4

Qtr3 15 24 17 17

Qtr4 18 15 23 23

>

# 19.8. Data Import

## 19.8.1. Quantitative Data

The quantitative data can be analysed in the following ways.

1. Frequency distribution of quantitative data
2. Histogram
3. Relative frequency distribution of quantitative data
4. Cumulative frequency distribution
5. Cumulative relative frequency distribution
6. Cumulative relative frequency graph

## 19.8.2. Frequency Distribution of Quantitative Data

This section gives the frequency distribution of ungrouped data and grouped data.

### 19.8.2.1. COUNT Function for Frequency of Ungrouped Data

The frequency of a given set of data can be obtained using count function in R. In the R program, the data are defined as vector and then library (plyr) should be invoked.

The library plyr can be installed using the following command at the prompt level of R.

>install.packages(plyr)

Then the command COUNT(name of the vector) gives the frequency distribution of the given data.

The required R program to find the frequency distribution of a set of sales data is given below.

> sales = c(20, 10, 12, 10, 20, 25, 10, 15, 30, 22, 34, 35, 22, 23, 39, 35, 40, 35)

> library(plyr)

> COUNT(sales)

x freq

1 10 3

2 12 1

3 15 1

4 20 2

5 22 2

6 23 1

7 25 1

8 30 1

9 34 1

10 35 3

11 39 1

12 40 1

>

### 19.8.2.2. COUNT Function for Frequency of Grouped Data

Frequency distribution of quantitative data gives the number of occurrences of the data in each interval of the data over the range of the data.

The frequency distribution of the sales figure under the salesman in all the four quarters is constructed in the following R program.

> sm1 = c(20, 22, 15, 18) # Sales data vector of salesman 1

> sm2 = c(22, 18, 24, 15) # Sales data vector of salesman 2

> sm3 = c(30, 25, 17, 23) # Sales data vector of salesman 3

> sm4 = c(25, 28, 20, 23) # Sales data vector of salesman 4

> s = data.frame(sm1, sm2, sm3, sm3) # data frame *s* containing all four vectors

> dimnames(s) = data.frame(c("Qtr1", "Qtr2", "Qtr3", "Qtr4"), c("SM1", "SM2", "SM3", "SM4")) #Names

> head(s) # Display of data frame *s* using head function

SM1 SM2 SM3 SM4

Qtr1 20 22 30 30

Qtr2 22 18 25 25

Qtr3 15 24 17 17

Qtr4 18 15 23 23

> salesfig = s$SM1 #sales figure of salesman 1 in all quarters

> range(salesfig) # Displaying range of data of vector sm1

[1] 15 22

>

> breaks = seq(15, 20, by = 5) # Forming breaks of frequency distribution

> breaks # Function to display the break points

[1] 15 20

> salesfig.cut = cut(salesfig, breaks, right = FALSE) # Function for sub-intervals of Freq. dist.

> salesfreq = table(salesfig.cut) # defining function for frequency distribution

> salesfig.freq # Function to display frequency distribution

salesfig.cut

(15, 20] (20, 25]

2 1

>

In the above R program, the steps of constructing the frequency distribution are listed below.

1. The vector sm1 named SM1 is defined as variable salesfig.
2. The range of the data under consideration for SM1 is displayed using range function.
3. The break points of the sub-intervals of the frequency distribution are defined and displayed using the function breaks.
4. Classify the data according to the breaks interval. The left-hand side of each sub-interval is defined, and the right-hand side of each sub-interval is left open.
5. Define the function salesfig.freq for the frequency of the data using table (salesfig.cut).
6. Display the frequency table using salesfig.freq.

The frequency table can be obtained column-wise using cbind function as shown in the following R program.

> cbind(salesfig.freq)

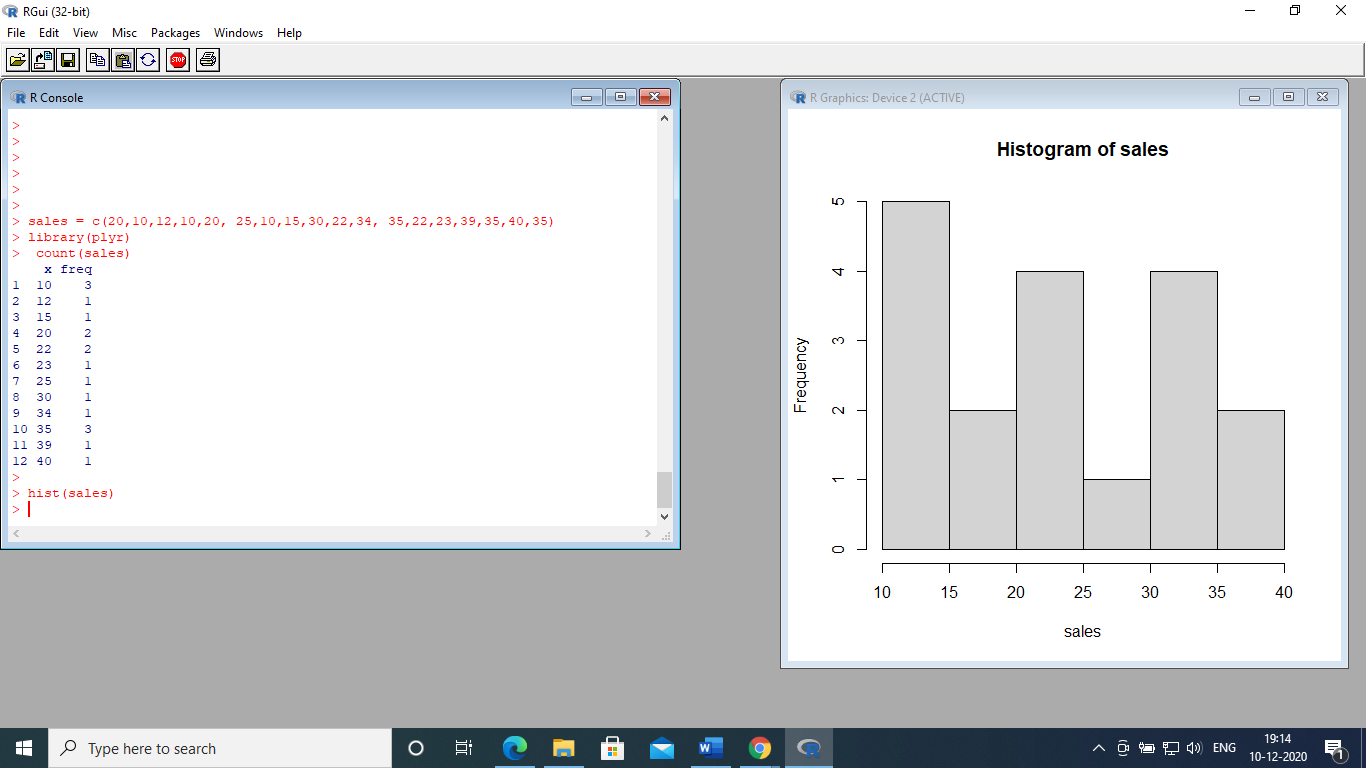
salesfig.freq

[15, 20) 2

[20, 25) 2

## 19.8.3. Histogram

The histogram of a given set of data is obtained using hist(name of the vector). The histogram of the sales data which has been shown in Section 19.8.1.1 is shown in Figure 19.6 using the hist() function.



**Figure 19.6.** Histogram of Sales Data Shown in Section 19.8.1.1

## 19.8.4. Cumulative Frequency Distribution

The summed-up frequencies of a frequency distribution form its cumulative frequency distribution. This section presents the construction of a cumulative distribution for ungrouped data as well as that for grouped data.

### 19.8.4.1. Cumulative Frequency Distribution of Ungrouped Data

The cumulative distribution of ungrouped data of weights of 12 students in a class is given by the following R program.

> weight = c(50, 52, 54, 52, 60, 58, 54, 50, 60, 52, 60, 52)

> weight.freq = table(weight)

> weight.freq = cumsum(weight.freq)

> weight.freq

50 52 54 58 60

2 6 8 9 12

>

### 19.8.4.2. Cumulative Frequency Distribution of Grouped Data

The cumulative frequency distribution of grouped data for the sales data of vector *sm1* in data frame *s* is given by the following R program.

> sm1 = c(20, 22, 15, 18)# Sales data vector of salesman 1

> sm2 = c(22, 18, 24, 15)# Sales data vector of salesman 2

> sm3 = c(30, 25, 17, 23)# Sales data vector of salesman 3

> sm4 = c(25, 28, 20, 23)# Sales data vector of salesman 4

> s = data.frame(sm1, sm2, sm3, sm3)# data frame *s* containing all four vectors

> dimnames(s) = data.frame(c("Qtr1", "Qtr2", "Qtr3", "Qtr4"), c("SM1", "SM2", "SM3", "SM4")) #Names

> head(s)

SM1 SM2 SM3 SM4

Qtr1 20 22 30 30

Qtr2 22 18 25 25

Qtr3 15 24 17 17

Qtr4 18 15 23 23

> salesfig = s$SM1 #sales figure of salesman 1 in all quarters

> breaks = seq(15, 25, by = 5) # Forming breaks of frequency distribution

> salesfig.cut = cut(salesfig, breaks, right = FALSE) # Function for subintervals of Freq. dist.

> salesfreq = table(salesfig.cut) # defining function for frequency distribution

> salesfreq.cumfreq = cumsum(salesfreq) # Defining cumulative frequency function

> salesfreq.cumfreq # Printing cumulative frequency for SM1

[15, 20) [20, 25)

2 4

**>**

## 19.8.5. Cumulative Frequency Graph

This section gives cumulative frequency graph of given data. A set of 18 sales values is considered as the sales data. The R program to construct the cumulative frequency for these data is given as follows. The cumulative frequency graph of the sales data obtained by the side of the R program in the same screen is shown as in Figure 19.7.

> sales = c(20, 10, 12, 10, 20, 25, 10, 15, 30, 22, 34, 35, 22, 23, 39, 35, 40, 35)

> breaks = seq(10, 40 by = 5)

> salesfig.cut = cut(salesfig, breaks, right = FALSE)

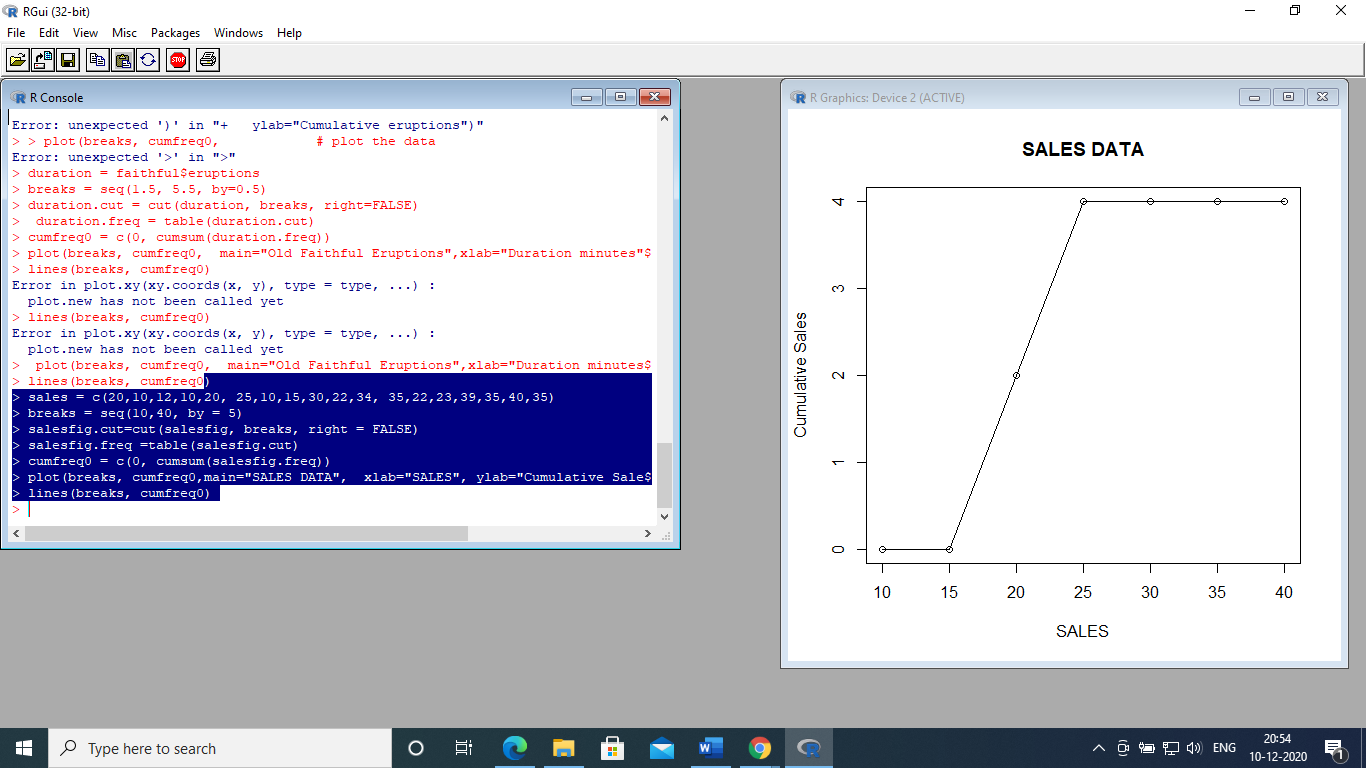
> salesfig.freq = table(salesfig.cut)

> cumfreq0 = c(0, cumsum(salesfig.freq))

> plot(breaks, cumfreq0,main ="SALES DATA", xlab ="SALES", ylab="Cumulative Sales")

+ # xlab is *X*-axis label and ylab is *Y*-axis label.

> lines(breaks, cumfreq0)



**Figure 19.7.** Cumulative Frequency Graph of Sales Data

# 19.9. Numerical Measures

This section presents different numerical measures of statistics, which are listed below.

* Mean
* Median
* Quartile
* Percentile
* Range
* Interquartile range
* Box plot
* Variance
* Standard deviation
* Covariance
* Correlation coefficient
* Skewness
* Kurtosis

## 19.9.1. Mean

The mean of a set of observations is the average of them, which has been already introduced in Excel.

An R program to obtain the mean of a given set of observations is given below.

sales = c(20, 10, 12, 10, 20, 25, 10, 15, 30, 22, 34, 35, 22, 23, 39, 35, 40, 35)

> mean(sales)

[1] 24.27778

>

The above R program defines the sales data in vector c, which in turn is stored in variables sales.

Next, the mean of these observations is obtained using the function mean(sales).

## 19.9.2. Median

Median is the middle most vale of a set of observations, when the observations are arranged in ascending order.

An R program to obtain the median of a given set of observations is given below.

> sales = c(20, 10, 12, 10, 20, 25, 10, 15, 30, 22, 34, 35, 22, 23, 39, 35, 40, 35)

> median(sales)

[1] 22.5

>

## 19.9.3. Quartile

Quartile is the value of a variable of interest corresponding to a percentage of the total frequency in steps of 25 per cent of the total frequency. The least quartile or the first quartile is the value of the variable with respect to 25 per cent of the total frequency. The second, third and fourth quartiles give the value of the variable of interest with respect to 50 per cent, 75 per cent and 100 per cent of the total frequency.

An R program to find the four quartiles, namely first quartile, second quartile, third quartile and fourth quartile of the sales data, is shown below. The function for the quartile is quantile (sales).

> sales = c(20, 10, 12, 10, 20, 25, 10, 15, 30, 22, 34, 35, 22, 23, 39, 35, 40, 35)

> median(sales)

[1] 22.5

> quantile(sales)

0 per cent 25 per cent 50 per cent 75 per cent 100 per cent

10.00 16.25 22.50 34.75 40.00

>

## 19.9.4. Percentile

Percentile is the value of the variable of concern say sales with respect to a percentage from 0 per cent to a given percentage of the total frequency of the observations of that variable, when they are arranged in ascending order.

An R program to find the percentile values of a variable say sales for 20 per cent of total frequency, 40 per cent of the total frequency and 73 per cent of the total frequency is shown below.

> sales = c(20, 10, 12, 10, 20, 25, 10, 15, 30, 22, 34, 35, 22, 23, 39, 35, 40, 35)

> quantile(sales, c(0.20, 0.40, 0.73))

20 per cent 40 per cent 73 per cent

13.20 21.60 34.41

>

The values of percentiles for 20 per cent, 40 per cent and 73 per cent are 13.2, 21.6 and 34.41, respectively.

## 19.9.5. Range

The range of a given data specifies the difference between the maximum value of the observations and the minimum value of the observations. This is explained using the sales data as shown in the following R program.

> sales = c(20, 10, 12, 10, 20, 25, 10, 15, 30, 22, 34, 35, 22, 23, 39, 35, 40, 35)

> max(sales) – min(sales)

[1] 30

>

The value of the range of the given set of observations is 30.

## 19.9.6. Interquartile Range

Interquartile range is the difference between the values of the variable corresponding to upper quartile and lower quartile of the observations.

The formula for the interquartile of a given set of observations is as given below.

Interquartile = Upper quartile – Lower quartile

An R program to find the interquartile of a given set of observations is given below.

> sales = c(20, 10, 12, 10, 20, 25, 10, 15, 30, 22, 34, 35, 22, 23, 39, 35, 40, 35)

> IQR(sales)

[1] 18.5

The value of the interquartile is 18.5.

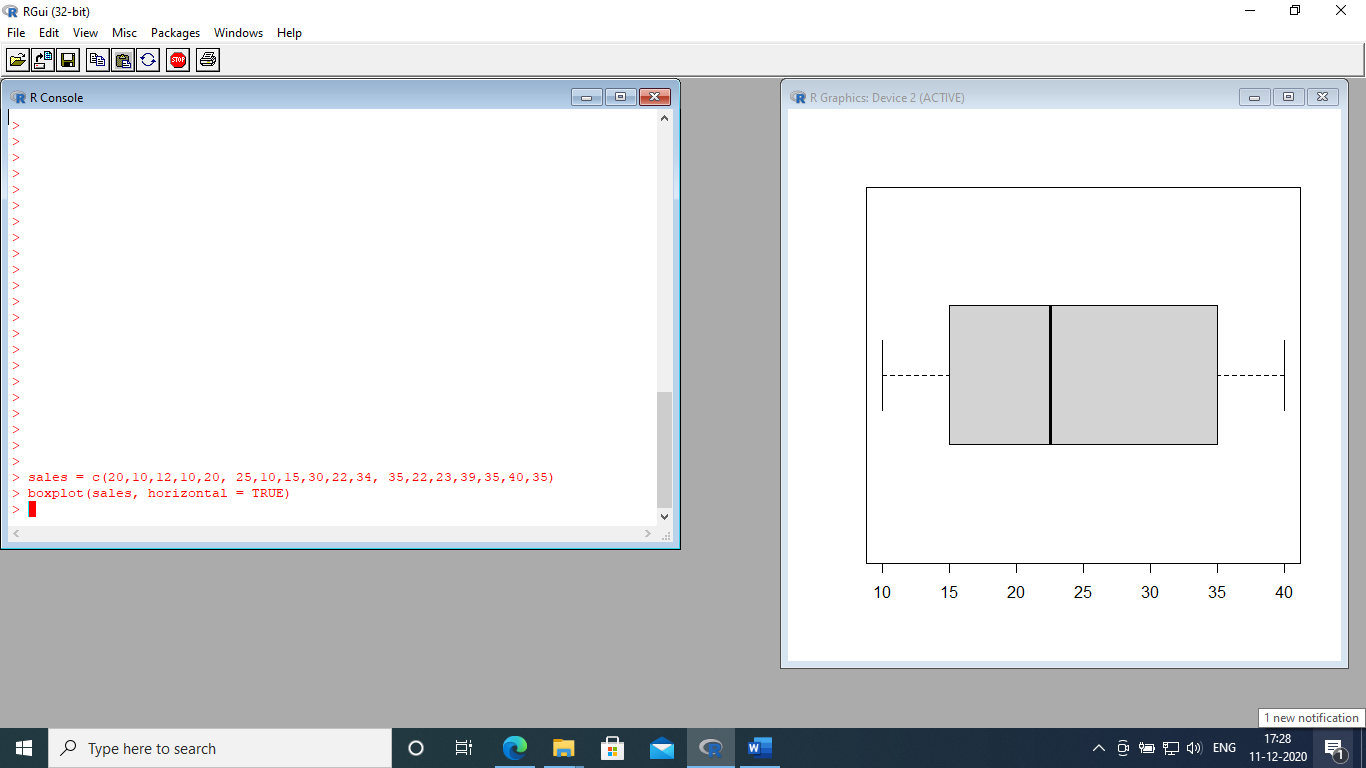
## 19.9.7. Box Plot

The box plot is based on quartiles and gives the lowest and highest values of the observations of a given set of observations. The visual presentation of this plot gives quick grasp of the shape of the distribution. An R program to get the box plot of the sales data is given below. The corresponding box plot is shown in Figure 19.8.

sales = c(20, 10, 12, 10, 20, 25, 10, 15, 30, 22, 34, 35, 22, 23, 39, 35, 40, 35)

> boxplot(sales, horizontal = TRUE)

>



**Figure 19.8.** Box Plot of Sales Data

## 19.9.8. Variance

The variance of a set of observations gives the dispersion of the observations around the mean of those observations. An R program to find the variance of the given set of data is shown below.

> sales = c(20, 10, 12, 10, 20, 25, 10, 15, 30, 22, 34, 35, 22, 23, 39, 35, 40, 35)

> var(sales)

[1] 107.8595

>

The variance of the sales data is 107.8595.

## 19.9.9. Standard Deviation

The standard deviation of a set of observation is the square root of the variance of those observations. An R program to find the variance of the given set of data is shown below.

> sales = c(20, 10, 12, 10, 20, 25, 10, 15, 30, 22, 34, 35, 22, 23, 39, 35, 40, 35)

> sd(sales)

[1] 10.38554

>

The standard deviation of the given set of observations of the sales is 10.38554.

## 19.9.10. Covariance

The covariance of two different streams of data explains the linear dependency among them. The formula of the covariance has been already given in Excel.

An R program to compute the covariance of the sales data and the advertising expenditure is shown below.

> sales = c(20, 10, 12, 10, 20, 25, 10, 15, 30, 22, 34, 35, 22, 23, 39, 35, 40, 35)

> adv\_expenditure = c(2, 1, 1.2, 1.1, 2.3, 2.6, 0.8, 1.9, 3.5, 2.5, 4, 4, 2, 2, 5, 2, 4, 5)

> cov(sales, adv\_expenditure)

[1] 12.36895

>

The covariance of the sales data and the advertising expenditure is 12.36895.

## 19.9.11. Correlation Coefficient

The correlation coefficient is the ratio of the covariance and the product of the individual standard deviations. This is a normalized measure of the linear dependency of the two variables.

An R program to compute the correlation coefficient of the sales data and the advertising expenditure is shown below.

> sales = c(20, 10, 12, 10, 20, 25, 10, 15, 30, 22, 34, 35, 22, 23, 39, 35, 40, 35)

> adv\_expenditure = c(2, 1, 1.2, 1.1, 2.3, 2.6, 0.8, 1.9, 3.5, 2.5, 4, 4, 2, 2, 5, 2, 4, 5)

> cor(sales, adv\_expenditure)

[1] 0.8928317

>

The value of the correlation coefficient among the variables, namely sales and adv\_expenditure, is 0.8928317.

## 19.9.12. Skewness

The skewness of a given set of data is a measure of skewness (distortion) of the distribution of the data as explained in Excel.

An R program to compute the skewness of the sales data is given below.

The package e1071 must be installed before running this program using ">install.packages("e1071")’.

> library(e1071)

> sales = c(20, 10, 12, 10, 20, 25, 10, 15, 30, 22, 34, 35, 22, 23, 39, 35, 40, 35)

> skewness(sales)

[1] 0.01580129

>

The skewness value is 0.01580129, which is very small and positive. Since the skewness is positive, the distribution is skewed towards right slightly.

## 19.9.13. Kurtosis

The excess kurtosis is a measure of the tail-shaped distribution as explained in Excel. An R program to find the excess kurtosis of the sales data is shown below.

The package e1071 must be installed before running this program using ">install.packages("e1071")’.

> library(e1071)

> sales = c(20, 10, 12, 10, 20, 25, 10, 15, 30, 22, 34, 35, 22, 23, 39, 35, 40, 35)

> kurtosis(sales)

[1] –1.504016

>

The excess kurtosis of the sales data is –1.504016. Since it is negative, the sales data distribution is platykurtic. The histogram of the data is not bell-shaped.

# 19.10. Probability Distributions

A probability distribution gives a curve by keeping the value of a random variable on *X*-axis and the probability of occurrence of that value of the random variable on *Y*-axis. The different probability distributions and the computation of the probability of the occurrence of the value of the random variables and corresponding cumulative probability are demonstrated using R program. The details of the probability distributions are already given under Excel.

## 19.10.1. Binomial Distribution

The binomial distribution represents the outcome of *n* independent trials. There will be only two outcomes (success/failure) in each trial. The probability of success *p* is known.

A quality control assistant checks the production units from a machine. The probability that the piece selected will be good is 0.8. He selected 12 units. Using R,

1. Find the probability that the number of good pieces is exactly 10.
2. Find the probability that the number of good pieces is less than or equal to 10.
3. The R program to compute the probability of having exactly 10 good pieces is given below.

> dbinom(10, size = 12, prob = 0.8)

[1] 0.2834678

>

1. The R program to compute the probability of having less than or equal to 10 good pieces is given below.

> pbinom(10, size = 12, prob = 0.8)

[1] 0.7251221

>

## 19.10.2. Poisson Distribution

The Poisson distribution represents the probability of occurrence of a specified arrival rate of customers in a queueing system (*x*). The mean of the Poisson distribution is . The details are already given under Excel.

The arrival rate of vehicles at a petrol bunk follows Poisson distribution with a mean arrival rate of 10 per 15 minutes interval. Find the following:

1. Probability of exactly 3 customers arriving in 15-minute interval.
2. Probability of at most 3 customers arriving in 15-minute interval.
3. Probability of at least 4 customers arriving in 15-minute interval.
4. The R program to find the probability of exactly 3 customers arriving in 15-minute interval is given below.

> dpois(3, lambda = 10)

[1] 0.007566655

>

1. The R program to find probability of at most 3 customers arriving in 15-minute interval is given below.

> ppois(3, lambda =10)

[1] 0.01033605

>

1. The R program to find probability of at least 4 customers arriving in 15-minute interval is given below.

**Approach 1**

> 1-ppois(3, lambda=10)

[1] 0.9896639

>

**Approach 2**

> ppois(3, lambda = 10, lower = FALSE)

[1] 0.9896639

>

## 19.10.3. Exponential Distribution

The exponential distribution represents the values of the probabilities for different values of random variable *x*, where *x* may be the service time in queueing system. Readers are requested to refer to this distribution under Excel, which has been already presented.

In a mainframe computer centre, execution times of programs follow exponential distribution. The average execution time (1/*μ*) is 0.75 minute. Find the probability that the execution time is

1. Equal to 2 minutes.
2. Less than 1.5 minutes.
3. More than 1.25 minutes.
4. In between 1 minute and 1.75 minutes.
5. The R program to find probability that the execution time is equal to 2 minutes is shown below.

> dexp(2, rate = 1/0.75)

[1] 0.0926446

>

1. The R program to find probability that the execution time is less than 1.5 minutes is shown below.

> pexp(1.5, rate = 1/0.75)

[1] 0.8646647

>

1. The R program to find probability that the execution time is more than 1.25 minutes is shown below.

> pexp(1.25, rate = 1/0.75, lower = FALSE)

[1] 0.1888756

>

d. The R program to find probability that the execution time is in between 1 minute and 1.75 minutes is shown below.

> pexp(1.75, rate = 1/0.75) – pexp(1, rate = 1/0.75)

[1] 0.1666252

>

## 19.10.4. Uniform Distribution

Uniform distribution is a continuous distribution. It represents equal probability of occurrence for each value in between the lower limit and upper limit of the random variable.

The R program to generate 12 uniformly distributed random numbers in interval 0 to 100 is shown below.

> runif(12, min = 0, max = 100)

[1] 8.9586714 13.0319992 39.5077290 50.4696294 0.3468246 70.8511102

[7] 88.5639450 5.9971710 99.7113760 32.6376462 48.3049720 27.2660613

>

## 19.10.5. Normal Distribution

The normal distribution is a bell-shaped continuous distribution. The readers are directed to refer to this distribution under Excel, which has been already presented.

In a survey with a sample of 200 private clinics, the daily number of patients treated follows normal distribution with its mean and standard deviation as 75 and 15, respectively. Answer the following.

a. What is the probability that the daily number of patients treated is less than 80?

b. What is the probability that the daily number of patients treated is more than 60?

a. The R program to find the probability that the daily number of patients treated is less than 80 is shown below.

> pnorm(80, mean = 75, sd =15, lower.tail = TRUE)

[1] 0.6305587

>

b. The R program to find the probability that the daily number of patients treated is more than 80 is shown below.

> pnorm(60, mean = 75, sd = 15, lower.tail = FALSE)

[1] 0.8413447

>

## 19.10.6. Student *t* Distribution

The student *t* distribution is also a bell-shaped distribution like the normal distribution with *n* – 1 degrees of freedom (df), where *n* is the number of observations.

Find the value of the random variable of *t* distribution with respect to a cumulative probability of 0.8 from left with df of 20.

The required R program is shown below.

> qt(0.8, df = 20)

[1] 0.8599644

>

## 19.10.7. Chi-square Distribution

The chi-square distribution is the sum of *m* independent standard normal variables with *m* df.

The R program to find the value of the chi-square random variable when the cumulative probability of 0.8 from left with a df 15 is shown below.

> qchisq(0.8, df = 15)

[1] 19.31066

>

## 19.10.8. F Distribution

F distribution is the ratio of two chi-square distributions with n1 df and n2 df, respectively.

The R program to find the value of the F random variable, when the cumulative probability of 0.85 from left with df 5 and 12, respectively, is shown below.

> qf(0.85, df1 = 5, df2 = 12)

[1] 2.005758

>

# 19.11. Interval Estimation

Interval estimation is the process of determining the span of the random variable of a probability distribution for a given significance level.

## 19.11.1. Point Estimation of Population Mean

This exercise requires mass function, which may be installed using install.packages(("mass").

The point estimate of a population mean is the mean of the sample observations.

An R program to compute the point estimate of the sales data is shown below.

> sales = c(20, 10, 12, 10, 20, 25, 10, 15, 30, 22, 34, 35, 22, 23, 39, 35, 40, 35)

> mean(sales)

[1] 24.27778

>

If the data are stored in a vector/data frame and if it has missing data, the following program is used.

> library(MASS)

> sales = c(20, 10, 12, 10, 20, 25, 10, 15, 30, 22, 34, 35, 22, 23, 39, 35, 40, 35)

> mean(sales, na.rm = TRUE)

[1] 24.27778

>

The point estimate of the sales is `24.27778 crore.

## 19.11.2. Interval Estimation of Population Mean with Known Variance

The formula for the interval estimation (confidence interval) when the sample size is large is given below.

± *ZXα*/2 (*σ*/√*n*)

Normal distribution is used for this estimate.

Consider an example that a company manufactures spindles in a lathe, which follows normal distribution with a variance of 121 mm.

A random sample of 49 spindles has been taken and the mean diameter of these spindles is 36 mm. Find the confidence interval of the diameter of the spindle that is manufactured using the lathe using a significance level of 0.05. The standard deviation of the population is 11 mm. Since the interval has two tails, half of the significance level (0.025) is kept at each end.

> library(MASS) # Invoking mass library

> m = 36 # Sample mean

> sdev = 11 # Sample standard deviation

> ss = 49 # Sample size

> sl = 0.05 # Significance level

> sem = sdev/sqrt(ss) # Standard error of mean

> E = qnorm(0.975) × sem # half interval

> m + c(–E,E) # Lower end and upper end of the confidence interval

[1] 32.92006 39.07994

>

The interval estimate (confidence interval) of the sample mean is from 32.92006 to 39.07994.

## 19.11.3. Interval Estimation of Population Mean with Unknown Variance

The formula for the interval estimation of population mean with unknown variance is given below.

± *tα*/2 (*s*/√*n*)

*t* distribution is used for this estimate.

It is demonstrated with a set of sales data as shown in the following R program.

> library(MASS) #Invoking MASS library

> sales = c(20, 10, 12, 10, 20, 25, 10, 15, 30, 22, 34, 35, 22, 23, 39, 35, 40, 35) # Sales vector

> m = mean(sales) # Mean of sales data

> n = 18 # Sample size

> df = n – 1 # Degrees of freedom

> sdev = sd(sales) # Standard deviation of sales

> sem = sdev/sqrt(n) # Standard error of mean of sales

> E = qt(0.975, df = n – 1) × sem # Half width of interval estimate

> m + c(–E, E) # Printing interval estimate

[1] 19.11317 29.44239

>

## 19.11.4. Sample Size of Population Mean

The sample size of population mean is determined using the following formula.

n =

where,

n is the sample size

is the z value on the right tail w.r.t

is the standard deviation

E is the half width of the interval estimate

Find the sample size if the standard deviation of the population mean of sales data is 11 crore, and the half interval precision is 1.5 crore with a significance level of 0.1.

The required R program is shown below.

> zalphabytwo = qnorm(0.95) #

> sigma = 11 # Standard deviation

> E = 1.5 # Half interval

> n = zalphabytwo^2 × sigma^2/E^2 # formula for sample size

> n

[1] 145.4981

>

The sample size (*n*) = 145.4981 = 146 (approximated)

# 19.12. Hypotheses Testing

An assumption about a population is called hypothesis.The process of testing the significance of the hypothesis is called hypothesis testing. In this section, hypotheses to conduct testing for different cases are presented.

## 19.12.1. Hypothesis Test of Population Mean with Known Variance at Left Tail of Large Sample

The greater than or equal to null hypothesis for testing population mean with known variance is shown below.

:

The annuals sales of companies in an industrial estate follows normal distribution with a mean of `30 crore. Its variance is 40. A sample of 35 companies is selected and its mean is `32 crore. Test whether the mean annuals sales is more than or equal to `30 crore at a significance level of 0.05.

: 30

The *z* statistic is :

The required R program is shown below.

> # part 1

> pm = 30 # Population mean

> sm = 32 # Sample mean

> psd = 20 # Population standard deviation

> n = 30 # Sample size

> z = (sm – pm)/(psd/sqrt(n)) # *z* statistic

> z

[1] 0.5477226

>

># Part 2

> alpha = 0.05 # Significance level

> zalpha = qnorm(1 – alpha) # value

> zalpha

[1] 1.644854

>

The value of z computed is less than the value of . Hence, reject , which means that the mean annual sales of the companies in the industrial estate is less than `30 crore.

## 19.12.2. Hypothesis Test of Population Mean with Known Variance at Right Tail of Large Sample

The null hypothesis for testing population mean with known variance and significance level placed at the right tail is shown below.

:

The annuals sales of companies in an industrial state follows normal distribution with a mean of `50 crore. Its variance is 90. A sample of 40 companies is selected and its mean is `49 crore. Test whether the mean annuals sales is less than or equal to `50 crore at a significance level of 0.05.

: 50

The *z* statistic is :

The required R program is shown below.

> Part 1

> pm = 50 # Population mean

> sm = 49 # Sample mean

> psd = 30 # Population standard deviation

> n = 40 # Sample size

> z = (sm – pm)/(psd/sqrt(n)) # *z* statistic

> z

[1] –0.2108185

> Part 2

> alpha = 0.05 # Significance level

> zalpha = qnorm(1 – alpha) # value

> zalpha

[1] 1.644854

>

The value of *z* computed is less than the value of . Hence, accept , which means that the mean sales of the companies in the industrial estate is more than `50 crore.

## 19.12.3. Two-tailed Test for Population Mean with Known Variance of Large Sample

The null hypothesis for two-tailed test for population mean with known variance is shown below.

:

The annuals sales of companies in an industrial state follows normal distribution with a mean of `100 crore. Its variance is 900. A sample of 36 companies is selected and its mean is `105 crore. Test whether the mean annuals is not different from `100 crore at a significance level of 0.05.

The *z* statistic is

: 100

The required R program is shown below.

> #Part 1

> pm = 100 # Population mean

> sm = 102 # Sample mean

> psd = 900 # Population standard deviation

> n = 36 # Sample size

> z = (sm – pm)/(psd/sqrt(n)) # *z* statistic

> z

[1] 0.01333333

>#Part 2

> alpha = 0.05 # Significance level

> zalpha = qnorm(1 – alpha/2) # Value of

> zalpha

[1] 1.959964

>

The acceptance range is from –1.959964 to 1.959964.

Since the value of *z* is in the acceptance region, accept the null hypothesis, which means that the mean sales of the companies does not differ from `100 crore.

## 19.12.4. Hypothesis Testing of Population Mean with Unknown Variance at Left Tail with Small Sample

The null hypothesis for testing of population mean with unknown variance at left tail of small sample is shown below.

:

The annuals sales of companies in an industrial state follows normal distribution with a mean of 25 crore. The sales in crores of rupees of 18 companies are 20, 10, 12, 10, 20, 25, 10, 15, 30, 22, 34, 35, 22, 23, 39, 35, 40 and 35. Test whether the mean annuals sales of the companies are more than or equal to 30 crore at a significance level of 0.05.

The required R program is shown below.

>#Part 1

> sales = c(20, 10, 12, 10, 20, 25, 10, 15, 30, 22, 34, 35, 22, 23, 39, 35, 40, 35) # Sales data

> meansales = mean(sales) # mean sales

> meansales

[1] 24.27778

> std.deviation = sd(sales) # Standard deviation of sales

> pop.mean = 25 # Population mean of sales

> n = 18 # Sample size

> t = (meansales – popmean)/(std.deviation/sqrt(n)) # *t* statistic

> t

[1] – 0.295038

> # part 2

> alpha = 0.05 #Significance level

> t.alpha = qt(1 – alpha, df = n – 1) # *t* statistic

> t.alpha

[1] 1.739607

>

Since the computed value of *t* is less than the cut-off value of *t*, reject the null hypothesis.

## 19.12.5. Hypothesis Test of Population Mean with Unknown Variance at Right Tail of Small Sample

The null hypothesis for testing population mean with unknown variance of small sample is shown below.

:

The annuals sales of companies in an industrial state follow normal distribution with a mean of `25 crore. A sample of 20 companies in an industrial estimate is taken and their mean sales is `27 crore and the standard deviation of the sales 12. Test whether the mean annuals sales is less than or equal to `25 crore at a significance level of 0.05.

The required R program is shown below.

>#Part 1

> popmean = 25 # Population mean

> sample.mean = 27 # Sample mean

> sigma = 12 # Sample standard deviation

> n = 20 # Sample size

> t = (popmean – sample.mean)/(sigma/sqrt(n)) # *t* statistic

> t

[1] – 0.745356

>#Part 2

> alpha = 0.05 # Significance value

> t.alpha = qt(1 – alpha, df = n – 1) # t cut-off value

> t.alpha

[1] 1.729133

>

Since the computed t value is less than the cut of t value, t falls in the acceptance region. Hence, accept the null hypothesis, which means that the mean sales of the company is less than `25 crore.

# 19.12.6. Two-tailed Test for Population Mean with Unknown Variance of Small Sample

The null hypothesis for two-tailed test for population mean with unknown variance of small sample is shown below.

:

The annuals sales of companies in an industrial state follows normal distribution with a mean of `200 crore. Its variance is 3,600. Sample of 25 companies is selected and its mean is `205 crore. Test whether the mean annuals sales is not different from `200 crore at a significance level of 0.05.

:

The required R program is shown below.

>#Part 1

> popmean = 200 # Population mean

> sample.mean = 205 # Sample mean

> sigma = 3,600 # Sample standard deviation

> n = 25 # Sample size

> t = (popmean – sample.mean)/(sigma/sqrt(n)) # *t* statistic

> t

[1] –0.006944444

>#Part 2

> alpha = 0.05

> t.alpha = qt(1 – alpha/2, df = n – 1)

> t.alpha

[1] 2.063899

>

The acceptance region is from –2.063899 to 2.063899. Since the computed t value of –0.006944444 is in the acceptance region, accept the null hypothesis, which means that the mean sales of the companies do not differ from `200 crore.

### 19.12.6.1. Another Example with Another Approach

The monthly sales of retail shops in a city are 6, 7, 5, 9, 7, 6, 5, 9, 5 and 3.

Check whether the mean sales are equal to 7 lakh with a significance level of 0.05.

:

The required R program with t.test ( ) is shown below.

> sales1 = c(6, 7, 5, 9, 7, 6, 5, 9, 5, 3)

>t.test(sales1)

One-sample *t*-test

data: sales1

t = 10.463, df = 9, *p*-value = 2.451e-06

alternative hypothesis: true mean is not equal to 0

95 per cent confidence interval:

4.859567 7.540433

sample estimates:

mean of x

6.2

>

The *p*-value is 2.451e-06, which is less than 0.025 (. Hence, reject the null hypothesis, which means that the mean of the monthly sales of the retail shops is different from 7 lakh.

## 19.12.7. One-tailed Test of Hypothesis Concerning Difference between Two Means of Population with Known Variances at Left Tail

The one-tailed test concerning the difference between two means of population with known variances is presented in this section.

,

The variance of the combined , =

Where,

is the variance of population 1

is the variance of population 2

= =

The income of companies in two different industrial estates follows normal distribution. The variances of population 1 and population 2 are `80 crore and `90 crore, respectively. The mean income of 36 companies in industrial estate 1 is `103 crore and that of 49 companies in industrial estate 2 is `108 crore, respectively. Check whether the difference between the means of two samples is more than 0 (*mean sales of sample 1 is more than that of sample 2*) with a significance level of 0.05.

The required R program is shown below.

>Part 1

> var1 = 80 # Variance of population 1

> var2 = 90 # Variance of population 2

> ms1 = 103 # Mean of sample 1

> ms2 = 108 # Mean of sample 2

> n1 = 36 # Size of sample 1

> n2 = 49 # Size of sample 2

> sig.combined = sqrt(var1^2/n1 + var2^2/n2) # Combined standard deviation

> z = (ms1 – ms2)/sig.combined # value of

> z

[1] –0.2699416

> #Part 2

> alpha = 0.05 # Significance level

> zalpha = qnorm(1 – alpha) # Cut-off value of at left tail

> zalpha

[1] 1.644854

>

The is less than the left cut-off value of , which lies in the rejection region. Hence, reject the null hypothesis, which means that the mean income of sample 1 is not greater than that of sample 2.

## 19.12.8. One-tailed Test of Hypothesis Concerning Difference between Two Means of Population with Known Variances at Right Tail

The one-tailed test concerning the difference between two means of population with known variances with the significance level kept at the right tail is presented in this section.

,

The variance of the combined , =

Where,

is the variance of population 1

is the variance of population 2

= =

The income of companies in two different industrial estates follows normal distribution. The variances of population 1 and population 2 are `100 crore and `90 crore, respectively. The mean income of 49 companies in industrial estate 1 is `100 crore and that of 36 companies in industrial estate 2 is 108 crore, respectively. Check whether the difference between the means of two samples is less than 0 with a significance level of 0.05.

The required R program is shown below.

>Part 1

> var1 = 100 # Variance of population 1

> var2 = 90 # Variance of population 2

> ms1 = 100 # Mean of sample 1

> ms2 = 108 # Mean of sample 2

> n1 = 49 # Size of sample 1

> n2 = 36 # Size of sample 2

> sig.combined = sqrt(var1^2/n1 + var2^2/n2) # Combined standard deviation

> z = (ms1 – ms2)/sig.combined # value of

> z

z

[1] –0.3862069

>

> #Part 2

> alpha = 0.05 # Significance level

> zalpha = qnorm(1 – alpha) # Cut-off value of at left

> zalpha

[1] 1.644854

>

The is less than the right cut-off value of , which lies in the acceptance region. Hence, accept the null hypothesis, which means that the mean income of sample 1 is less than that of sample 2.

## 19.12.9. Two-tailed Test of Hypothesis Concerning Difference between Two Means of Population with Known Variances

The two-tailed test concerning the difference between two means of population with known variances is presented in this section.

,

The standard deviation of the combined variance, =

Where,

is the variance of population 1

is the variance of population 2

= =

The income of companies in two different industrial estates follows normal distribution. The variances of population 1 and population 2 are `50 crore and `55 crore, respectively. The mean income of 36 companies in industrial estate 1 is `80 crore and that of 49 companies in industrial estate 2 is `90 crore. Check whether the difference between the means of two samples does not differ from 0 with a significance level of 0.05.

The required R program is shown below.

The required R program is shown below.

>Part 1

> var1 = 50 # Variance of population 1

> var2 = 55 # Variance of population 2

> ms1 = 80 # Mean of sample 1

> ms2 = 90 # Mean of sample 2

> n1 = 36 # Size of sample 1

> n2 = 49 # Size of sample 2

> sig.combined = sqrt(var1^2/n1 + var2^2/n2) # Combined standard deviation

> z = (ms1 – ms2)/sig.combined # value of

> z

[1] –0.8731073

>

> #Part 2

> alpha = 0.05 # Significance level

> zalpha = qnorm(1 – alpha) # Cut-off value of at left

> zalpha

[1] 1.644854

>

The acceptance range of the mean income is from –1.644854 to 1.644854.

The lies in the acceptance region. Hence, accept the null hypothesis, which means that the mean income of population 1 does not differ from that of population 2.

## 19.12.10. Two-tailed Test of Hypothesis Concerning Difference between Two Means of Population with Unknown Variances

Let *X*1 and *X*2 be two random variables with respect to two independent populations and follow normal distribution. Assume that their variances are unknown and equal. The sample size of each population is less than or equal to 30.

In this situation, *σ*12 = *σ*22

Type 3 test: (Both tails)

*H*0: *μ*1 – *μ*2 = 0

*H*1: *μ*1 – *μ*2 ≠ 0

In this type of hypothesis testing, half of the significant level (α/2) is placed at both tails of the distribution.

The standard deviation of the difference in the population means cannot be estimated using the sample standard deviations *S*1 and *S*2. Hence, a polled variance is estimated using the following formula with the assumption that the variances of the samples are equal (*σ*2 = *σ*12 = *σ*22).

The formula of the standard deviation for the difference between population means is given below.

The *t* statistic for this reality is given below.

with (*n*1 + *n*2 – 2) df.

The weekly sales in lakhs of rupees of 10 retail shops in City 1 and those of 10 retail shops in City 2 are shown below.

City 1: (6, 7, 5, 9, 7, 6, 5, 9, 5, 3)

City 2: (8, 4, 5, 6, 3, 2, 5, 6, 2, 4)

Check whether the mean sales of the retail shops in city 1 does not differ from that of the retail shops in city 2 with a significance level of 0.05.

> sales1 = c(6, 7, 5, 9, 7, 6, 5, 9, 5, 3)

> sales2 = c(8, 4, 5, 6, 3, 2, 5, 6, 2, 4)

> t.test(sales1, sales2)

Welch two-sample *t*-test

data: sales1 and sales2

t = 2.0144, df = 17.996, *p*-value = 0.05916

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

95 per cent confidence interval:

–0.07306164 3.47306164

sample estimates:

mean of x mean of y

6.2 4.5

>

The value of *p* is 0.05916, which is more than 0.025 (. Hence, accept the null hypothesis, which means that the mean monthly sales of the retail shops in City 1 is same as that of the retail shops in City 2.

# 19.13. Chi-square Test

This section presents chi-square test applied to the following cases.

* Chi-square test for independence
* Goodness-of-fit test

## 19.13.1 Chi-square Test for Independence

The chi-square test for independence is used to check whether there is dependency between the levels of two categories of data.

The hypotheses of this test are as given below.

*H*0: *p*ij = *p*i. X *p*.j , i = 1, 2, 3, …, *m* and j = 1, 2, 3, …, *n*

This means that the levels of category *A* and the levels of category *B* are independent in terms of their frequencies.

*H*1: *p*ij > *p*i. × *p*.j for at least one combination of i and j,

where i = 1, 2, 3, …, *m* and j = 1, 2, 3, …, *n*.

This means that the levels of category *A* and the levels of category *B* are not independent in terms of their frequencies.

Consider the contingency table for salesmen, SM1, SM2, SM3 and SM4 under the category called salesman and the quarters, Qtr1, Qtr2, Qtr3 and Qtr4 under the category quarter as shown in Table 19.2.

**Table 19.2.** Contingency Table of Salesman and Quarter for Sales in Crores

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Salesman | | | |
|  | SM1 | SM2 | SM3 | SM4 |
| Qtr1 | 20 | 22 | 30 | 25 |
| Qtr2 | 22 | 18 | 25 | 28 |
| Qtr3 | 15 | 24 | 17 | 20 |
| Qtr4 | 18 | 15 | 23 | 23 |

The levels of the two categories, namely salesman and quarter, are independent in terms of sales.

Check whether there is dependency among the levels of the categories, namely salesman and quarter, in terms of sales at a significance level of 0.05.

The required R program is shown below.

> sm1 = c(20, 22, 15, 18)

> sm2 = c(22, 18, 24, 15)

> sm3 = c(30, 25, 17, 23)

> sm4 = c(25, 28, 20, 23)

> s = data.frame(sm1, sm2, sm3, sm4)

> dimnames(s) = data.frame(c("Qtr1","Qtr2","Qtr3","Qtr4"), c("SM1","SM2","SM3","SM4"))

> s

SM1 SM2 SM3 SM4

Qtr1 20 22 30 25

Qtr2 22 18 25 28

Qtr3 15 24 17 20

Qtr4 18 15 23 23

> chisq.test(s)

Pearson’s Chi-squared test

data: s

*X*-squared = 5.5922, df = 9, *p*-value = 0.7799

>

The *p*-value is 0.7799, which is more than the given significance level of 0.05. Hence, accept the null hypothesis. This means that the levels of the two categories, salesman and quarter, are independent in terms of sales.

## 19.13.2. Goodness-of-fit Test

The given data may follow a certain pattern. Through scatter plot, one can find out the shape of the distribution to which the given data resemble. Then, the goodness-of-fit test can be carried out to confirm the assumed distribution at a significance level using chi-square test.

The null and alternate hypotheses for this goodness-of-fit test are given below.

**H~~0~~:** The given data follow an assumed probability distribution.

**H1:** The given data do not follow the assumed probability distribution.

The daily demand of a product follows uniform distribution. The observed frequency of the demand values is summarized in Table 19.3.

**Table 19.3.** Observed Frequencies of Demand

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Demand | 26 | 27 | 28 | 29 | 30 | 31 | 32 | 33 | 34 | 35 |
| Observed frequency (*o*i) | 17 | 15 | 13 | 14 | 11 | 15 | 17 | 14 | 19 | 15 |

Check whether the given data follow uniform distribution at a significance level of 0.05.

The required R program is given below.

> observed.freq = c(17, 15, 13, 14, 11, 15, 17, 14, 19, 15)

> observed.freq

[1] 17 15 13 14 11 15 17 14 19 15

> chisq.test(observed.freq)

Chi-squared test for given probabilities

data: observed.freq

*X*-squared = 3.0667, df = 9, *p*-value = 0.9616

>

The *p*-value is 0.9616, which is more than the given significance level of 0.05. Hence, accept the null hypothesis, which means that the given data follow uniform distribution.

# 19.14. Analysis of Variance (ANOVA)

Analysis of variance aims to test the effect of factor(s) on the response variable of interest.

This section presents ANOVA for the following designs.

* Completely randomized design
* Randomized block design
* Factorial design

## 19.14.1. Completely Randomized Design

The completely randomized design considers only one factor with *k* treatments and *n* replications in each treatment.

**:** There are no significant differences among the treatments of the factor in terms of the response variable.

**:** There are significant differences among the treatments of the factor in terms of the response variable.

A stock market analyst wants to study the effect of the type of the company on the earnings per share (*EPS*). So he collected *EPS* data for the past six years of three different types of company from a secondary source, which are summarized in Table 19.4. Check whether there are significant differences among the companies in terms of yearly *EPS* data at a significance level of 0.05.

Table 19.4. Yearly EPS Data of Companies

|  |  |  |  |
| --- | --- | --- | --- |
| Replication | Company | | |
| 1 | 2 | 3 |
| 22 | 52 | 16 |
| 42 | 33 | 24 |
| 44 | 8 | 19 |
| 52 | 47 | 18 |
| 45 | 43 | 34 |
| 37 | 32 | 39 |

The required R program is given below.

> r = c(22, 42, 44, 52, 45, 37, 52, 33, 8, 47, 43, 32, 16, 24, 19, 18, 34, 39) # Replications of treatments in sequence

> r

[1] 22 42 44 52 45 37 52 33 8 47 43 32 16 24 19 18 34 39

> f = c("C1","C2","C3") # Assigning name for treatments of factor company

> k = 3 # Number of treatments of factor company

> n = 6 # Number of replications under each treatment

> tm = gl(k, 1, n × k, factor(f)) # Matching replications to treatments of factor

> tm

[1] C1 C2 C3 C1 C2 C3 C1 C2 C3 C1 C2 C3 C1 C2 C3 C1 C2 C3

Levels: C1 C2 C3

> avar = aov(r ~ tm) # Invoking formula for ANOVA

> summary(avar)

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

tm 2 152.4 76.22 0.409 0.671

Residuals 15 2,793.2 186.21

>

The *p*-value is 0.671, which is more than the given significance level of 0.05. Hence, accept the null hypothesis.

**Inference:** There are no significant differences among the companies in terms of EPS value.

## 19.14.2. Randomized Block Design

The randomized block design considers only one factor with *k* treatments and *n* replications in each treatment. The replications across the treatments are considered as blocks. First replications of all the treatments of the factor put together is called block 1 and so on.

### Factor

**:** There are no significant differences among the treatments of the factor in terms of the response variable.

**:** There are significant differences among the treatments of the factor in terms of the response variable.

### Block

**:** There are no significant differences among the blocks in terms of the response variable.

**:** There are significant differences among the blocks in terms of the response variable.

A stock market analyst wants to study the effect of the type of the company on the *EPS*. So he collected *EPS* data for the past six years of three different types of company from a secondary source, which are summarized in Table 19.5. First replications of all the treatments of the factor put together is called block 1 and so on.

1. Check whether there are significant differences among the companies in terms of yearly *EPS* data at a significance level of 0.05.
2. Check whether there are significant differences among the blocks in terms of yearly *EPS* data at a significance level of 0.05.

**Table 19.5.** Yearly EPS Data of Companies

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Block |  | Company | | |
| 1 | 2 | 3 |
| 1 | 22 | 52 | 16 |
| 2 | 42 | 33 | 24 |
| 3 | 44 | 8 | 19 |
| 4 | 52 | 47 | 18 |
| 5 | 45 | 43 | 34 |
| 6 | 37 | 32 | 39 |

The required R program is given below.

> r = c(22, 42, 44, 52, 45, 37, 52, 33, 8, 47, 43, 32, 16, 24, 19, 18, 34, 39) # Replications of treatments in sequence

> r

[1] 22 42 44 52 45 37 52 33 8 47 43 32 16 24 19 18 34 39

> f = c("C1","C2","C3") # Assigning name for treatments of factor company

> k = 3 # Number of treatments of factor company

> n = 6 # Number of replications under each treatment

> tm = gl(k, 1, n × k, factor(f)) # Matching replications to treatments of factor

> tm

[1] C1 C2 C3 C1 C2 C3 C1 C2 C3 C1 C2 C3 C1 C2 C3 C1 C2 C3

Levels: C1 C2 C3

> block = gl(n, k, k × n)

> block

[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

> avar = aov(r ~ tm + block) # Invoking ANOVA for randomized block design

> summary(avar)

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

tm 2 152.4 76.22 0.469 0.639

block 5 1,168.9 233.79 1.439 0.291

Residuals 10 1,624.2 162.42

>

### Inferences

1. The *p*-value of treatment is 0.639, which is more than the given significance level of 0.05. Hence, accept the null hypothesis with respect to the treatment. This means that there are no significant differences among the companies in terms of EPS value.
2. The *p*-value of block is 0.291, which is more than the given significance level of 0.05. Hence, accept the null hypothesis with respect to the block. This means that there are no significant differences among the blocks in terms of EPS value.

## 19.14.3. Factorial Design

The factorial design considers more than one factor and tests the significance of each of the components of the ANOVA model at a given significance level. The readers are directed to refer to the corresponding chapter under Excel.

A stock market analyst wants to study the effect of the type of company as well as region and their interaction on the *EPS*. So he collected *EPS* data for the past three years of three different companies in each of the regions, namely North and South of a nation from secondary sources, which are summarized in Table 19.6. The yearly observations form the replications under each combination of company and region.

1. Check whether there are significant differences among the companies in terms of yearly *EPS* data at a significance level of 0.05.
2. Check whether there are significant differences among the regions in terms of yearly *EPS* data at a significance level of 0.05.
3. Check whether there are significant differences among the interactions in terms of company and region, and in terms of yearly *EPS* data at a significance level of 0.05.

**Table 19.6.** Yearly EPS Data with Respect to Companies and Regions

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Region |  |  | Company | | |
| 1 | 2 | 3 |
| North | 1 | 22 | 52 | 16 |
| 2 | 42 | 33 | 24 |
| 3 | 44 | 8 | 19 |
| South | 1 | 52 | 47 | 18 |
| 2 | 45 | 43 | 34 |
| 3 | 37 | 32 | 39 |

Check the significance of each of the components of the ANOVA model at a significance level of 0.05.

The hypotheses of the two factors and their interaction term are listed below.

### Factor: Company

**H0:** There are no significant differences between the treatments of company in terms of EPS.

**H1:** There are significant differences between the treatments of company in terms of EPS.

### Factor: Region

**H0:** There are no significant differences between the treatments of region in terms of EPS.

**H1:** There are significant differences between the treatments of region in terms of EPS.

### Interaction: Company × Region

**H0:** There are no significant differences between different pairs of company and region in terms of EPS.

**H1:** There are significant differences between different pairs of company and region in terms of the EPS.

The required R program is given below.

> r = c(22, 42, 44, 52, 45, 37, 52, 33, 8, 47, 43, 32, 16, 24, 19, 18, 34, 39) # Replications of treatments in sequence

> r

[1] 22 42 44 52 45 37 52 33 8 47 43 32 16 24 19 18 34 39

> fact1 = c("C1","C2","C3") # Defining the treatments of the factor company

> fact1

[1] "C1" "C2" "C3"

> fact2 = c("Estate 1", "Estate 2") # Defining the treatments of the factor region

> fact2

[1] "Estate 1" "Estate 2"

> k1 = length(fact1) # Defining the number of treatments of company

> k2 = length(fact2) # Defining the number of treatments of region

>n = 3 # Defining the number of replications

>tm1 = gl(k1, 1, n × k1 × k2, factor(fact1)) # Defining replications for treatments of company

> tm1

[1] C1 C2 C3 C1 C2 C3 C1 C2 C3 C1 C2 C3 C1 C2 C3 C1 C2 C3

Levels: C1 C2 C3

> tm2 = gl(k2, n × k1, n × k1 × k2, factor(fact2)) # Defining the replications for treatments   
 of region

> tm2

[1] North North North North North North North North North South South South

[13] South South South South South South

Levels: North South

> avar = aov(r ~ tm1 × tm2) # Invoking the formulas for factorial experiment

> summary(avar) # Summary of ANOVA table

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

tm1 2 152.4 76.22 0.382 0.691

tm2 1 220.5 220.50 1.105 0.314

tm1:tm2 2 177.3 88.67 0.444 0.651

Residuals 12 2,395.3 199.61

>

### Inferences

1. The *p*-value of company is 0.691, which is more than the given significance level of 0.05. Hence, accept the null hypothesis with respect to the treatments of company. This means that there are significant differences among the companies in terms of EPS.
2. The *p*-value of region is 0.314, which is more than the given significance level of 0.05. Hence, accept the null hypothesis with respect to the treatments of region. This means that there are significant differences among the regions in terms of EPS.
3. The *p*-value of interaction between company and region is 0.651, which is more than the given significance level of 0.05. Hence, accept the null hypothesis with respect to the interaction terms.

# 19.15. Non-parametric Test

As already stated, non-parametric tests do not use parameter(s) of the data in hypothesis testing. This section presents the following non-parametric tests.

* + Sign test
  + Mann–Whitney U test using Excel sheet
  + Kruskal–Wallis test (H test)

## 19.15.1. Sign Test

In one-tailed one-sample sign test, when sample size is small,a small random sample is taken from a non-normal population and then it is tested against a median value (*μ*) such that the observations in the sample are more than that median *(μ*) or less than that median *(μ*) at a significance level of *α* using Binomial distribution.

#### Example 11.1

A time study engineer collected data on the final assembly operation in an assembly line producing a two-wheeler. The number of observations is 9, namely 21, 34, 28, 15, 27, 15, 27, 26 and 12. Check the null hypothesis that whether the assembly operation time is 20 minutes (*H*o: *μ* = 20) as against the alternate hypothesis *H*1: *μ* > 20 using sign test as a significance level of 0.05.

Sample size (*n*) = 9

Significance level (*α* = 0.05)

The observations of the final assembly operations time: 21, 34, 28, 15, 27, 15, 27, 26 and 12.

Let *X* be a random variable representing plus sign when 20 is subtracted from each observation.

*H*o: *μ* = 20 or p = ½

*H*1: *μ* > 20 or p > ½

This is based on binomial test, which has the following syntax.

>binom.test(number of successes, total number of observations)

Number of success (when value = 3, n = 9

The required R program is shown given below.

> binom.test(3, 9)

Exact binomial test

data: 3 and 9

number of successes = 3, number of trials = 9, *p*-value = 0.5078

alternative hypothesis: true probability of success is not equal to 0.5

95 per cent confidence interval:

0.07485463 0.70070494

sample estimates:

probability of success

0.3333333

>

Since the *p*-value 0.3333333 is more than the given significance level, accept the null hypothesis. This means that the assembly operation time is 20 min.

## 19.15.2. Mann–Whitney U Test Using Excel Sheet

Mann–Whitney U test is an alternate to two samples *t* test and it is powerful. This test is based on the ranks of the combined observations of the samples. This is also called rank-sum test.

Let

*n*1 be the size of sample 1

*n*2 be the size of sample 2

*N* = *n*1 + *n*2

The null and alternate hypotheses of this test are listed below.

Test 1 (smaller than type)

*H*o: The two samples are drawn from different populations having the same distribution.

*H*1: Population 1 is stochastically lesser than population 2.

In a survey conducted by an investigator, the investigator has appointed two enumerators (enumerator 1 and enumerator 2). The number of respondents covered by these enumerators during randomly selected days are summarized in Table 19.7.

**Table 19.7.** Number of Respondents Covered by Enumerators

|  |  |  |
| --- | --- | --- |
| Day | Enumerator 1 | Enumerator 2 |
| 1 | 24 | 30 |
| 2 | 17 | 20 |
| 3 | 34 | 15 |
| 4 | 28 | 22 |
| 5 | 15 | 31 |
| 6 | 35 | 24 |
| 7 | 25 | 12 |
| 8 | 13 | 16 |
| 9 | 22 | 21 |
| 10 | 28 | 8 |
| 11 | 27 | 22 |
| 12 | 15 | 14 |
| 13 | 13 | 16 |
| 14 | 17 |  |

Check whether the two samples shown in Table 19.7 are drawn from identical populations against the alternate hypothesis that population 1 stochastically differs from population 2 using Mann–Whitney U test at a significance level of 0.01.

*n*1 be the size of sample 1

*n*2 be the size of sample 2

*N* = *n*1 + *n*2

*Ho*: The two samples are drawn from different populations having the same distribution.

*H*1: Population 1 stochastically differs from population 2.

The required R program is given below.

> enumerator1 = c(24, 17, 34, 28, 15, 35, 25, 13, 22, 28, 27, 15, 13, 17)

> enumerator2 = c(30, 20, 15, 22, 31, 24, 12, 16, 21, 8, 22, 14, 16)

> wilcox.test(enumerator1, enumerator2)

Wilcoxon rank-sum test with continuity correction

data: enumerator1 and enumerator2

W = 112.5, *p*-value = 0.3072

alternative hypothesis: true location shift is not equal to 0

Warning message:

In wilcox.test.default(enumerator1, enumerator2):

cannot compute exact *p*-value with ties

>

The value of *p* is 0.3072, which is more than the given significance level of 0.01. Hence, accept the null hypothesis, which is stated below.

*Ho*: The two samples are drawn from different populations having the same distribution.

## 19.15.3. Kruskal–Wallis Test (H Test) Using Excel Sheet

Kruskal–Wallis test is like the *K*-samples median test in which the objective is to test whether the *K* samples are from K identical populations. This test is alternatively called *H* test and the statistic that is computed for this test is denoted by *H*. This test is an alternate approach for ANOVA with single factor.

The hypotheses of the *H* test are listed below.

*H*o: *K* samples, which are independent, are drawn from *K* identical populations.

*H*1:*K* samples, which are independent, are not drawn from *K*identicalpopulations.

A stock market analyst wants to study the effect of the type of the company on the yearly *EPS*. So he collected *EPS* data for the past four years of four different companies from a secondary source, which are summarized in Table 19.8. Check whether there are significant differences among the companies in terms of yearly *EPS* data at a significance level of 0.05.

**Table 19.8.** Yearly EPS Data of Companies

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Replications | Company | | | |
| C1 | C2 | C3 | C4 |
| 12 | 16 | 9 | 13 |
| 8 | 18 | 22 | 8 |
| 15 | 11 | 15 | 5 |
| 18 | 10 | 25 | 20 |

The required R program is given below.

> c1 = c(12, 18, 15, 18)

> c1 = c(12, 8, 15, 18)

> c2 = c(16, 18, 11, 10)

> c3 = c(9, 22, 15, 25)

> c4 = c(13, 8, 5, 20)

> epsdata = data.frame(c1, c2, c3, c4)

> kruskal.test(epsdata)

Kruskal–Wallis rank-sum test

data: epsdata

Kruskal–Wallis chi-squared = 2.0938, df = 3, *p*-value = 0.5532

>

Since the *p*-value of 0.5532 is more than the given significance level of 0.05, accept the null hypothesis, which means that there are no significant differences among the companies in terms of yearly EPS value.

# 19.16. Regression Analysis

This section presents fitting of regression model for each of the following cases.

* Simple regression analysis
* Multiple regression analysis

## 19.16.1. Simple Regression Analysis

Simple regression aims to develop an equation, which contains an independent variable (*x*) and a dependent variable (*y*) as shown below.

*y* = *a + b x*

The equation can be used to predict the value of the dependent variable (*y*) for a given value of the independent variable (*x*).

This is based on least square principle.

Consider the demand of a product during the past 10 years as shown below.

Month 1 2 3 4 5 6 7 8 9 10 11 12

Demand

(units) 90 100 110 130 125 145 160 150 180 198 210 205

Fit a simple regression model to estimate the demand of the product for a given month in future.

The required R program is shown below.

>Part 1 fitting regression model to predict demand using month

> month = c(1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12) # Values of independent variable

> demand = c(90, 100, 110, 130, 125, 145, 160, 150, 180, 198, 210, 205) # Values of dependent  
 variable

> reg.model = lm(demand ~ month) # Fitting regression model

> print(reg.model) # Printing coefficients of model

Call:

lm(formula = demand ~ month)

Coefficients:

(Intercept) month

77.27 11.23

# Part 2

> print(summary(reg.model)) # printing summary of regression model

Call:

lm(formula = demand ~ month)

Residuals:

Min 1Q Median 3Q Max

–17.0909 –2.4659 0.9318 5.0568 9.2273

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 77.2727 4.9729 15.54 2.49e-08 \*\*\*

month 11.2273 0.6757 16.62 1.30e-08 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.08 on 10 df

Multiple *R*-squared: 0.965, adjusted *R*-squared: 0.9616

*F* statistic: 276.1 on 1 and 10 df, *p*-value: 1.303e-08

>

### 19.16.1.1. Regression Model and Checking Statistical Significance of Regression

From the result of part 1 of R program, the model is as given below.

demand = 77.2727 + 11.2273 month

The results of part 2 give the summary of the regression model.

The value of *p* is 1.303, which is less than the significance level of 0.05. Hence, the null hypothesis of regression is rejected. This means that the demand depends on the month with statistical significance, and it predicts well the demand of the product for a given month.

## 19.16.2. Multiple Regression Analysis

The multiple regression model aims to fit a regression model to estimate the value of the dependent variable in it for a given set of independent variables, which means that the number of independent variables is more than one.

* A generalized multiple regression model is shown below.

y =

where,

*n* is the number of independent variables

is the *i*thindependent variable, *i* = 1, 2, 3, …, *n*

*y* is the dependent variable

is the intercept

is the coefficient of the *i*th independent variable, *i* = 1, 2, 3, …, *n*

Consider the annual sales in crores of rupees of a product during the past 10 years as shown below, which are assumed to be affected by "sales force" and "advertising expenditure".

Annual sales

(Crore of Rs.) 20, 18, 30, 18, 19, 22, 20, 22, 20, 28

Sales force 8, 6, 10, 7, 8, 10, 9, 11, 10, 11

Annual advertising

Expenditure 28, 23, 38, 16, 20, 28, 23, 30, 26, 32

Fit a multiple regression model to estimate the annual sales of the product for a given sales force and advertising expenditure and test its significance at a significance level of 0.05.

y =

where,

is the independent variable representing "sales force"

is the independent variable representing "annual advertising expenditure"

*y* is the dependent variable, which is "annual sales"

is the intercept

is the coefficient of the independent variable "sales force"

is the coefficient of the independent variable "annual advertising expenditure"

The required R program is shown below.

>Part 1 Designing Multiple Regression Model

> x1 = c(8, 6, 10, 7, 8, 10, 9, 11, 10, 11)

> x2 = c(28, 23, 38, 16, 20, 28, 23, 30, 26, 32)

> y = c(20, 18, 30, 18, 19, 22, 20, 22, 20, 28)

> datasource = data.frame(c(y, x1, x2))

> mul.regmodel = lm(y ~ x1 + x2, datasource)

> mul.regmodel

Call:

lm(formula = y ~ x1 + x2, data = datasource)

Coefficients:

(Intercept) x1 x2

5.1676 0.3301 0.5137

>Part 2 Summary of Multiple Regression Model

print(summary(mul.regmodel))

Call:

lm(formula = y ~ x1 + x2, data = datasource)

Residuals:

Min 1Q Median 3Q Max

–2.2095 –1.6092 –0.4027 1.7376 2.7631

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 5.1676 3.8969 1.326 0.2264

x1 0.3301 0.5885 0.561 0.5924

x2 0.5137 0.1589 3.232 0.0144 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.155 on 7 df

Multiple *R*-squared: 0.7864, adjusted *R*-squared: 0.7253

*F* statistic: 12.88 on 2 and 7 df, *p*-value: 0.004507

>

### 19.16.2.1. Fitted Multiple Regression Model

Based on the result of Part 1, the multiple regression model obtained using R program is shown below.

y = 5.1676 +

### 19.16.2.2. Significance of Regression Coefficients

Based on the *p*-values of the regression model in part 2 of the R program, the following inferences are drawn.

* The *p*-value of the regression model is 0.004507, which is less than the assumed significance level of 0.05. Hence, the regression model is having best fit by rejecting the null hypothesis.
* Further, the *R*-squared value of the regression model is 0.7864, which is also good enough to support the above inference.

# Summary

* R programming is a programming software, which has the feature of having functions for most of the modules of computing, namely vector, matrix multiplication, measure of central tendencies, regression model and so on. When this is compared to any high-level language, the extent of code to be written is very minimal.
* The data types handled in R program are numeric, integer, complex, logic and character.
* A matrix is a two-dimensional array having a specified number of rows and a specified number of columns.
* A list contains more than one vector.
* The member of vectors included in a list can be accessed using double indexing.
* In R programming, data frame is used to store data table, which consists of a list of vectors of equal length.
* The frequency of a given set of data can be obtained using COUNT function in R.
* The summed-up frequencies of a frequency distribution form its cumulative frequency distribution.
* Frequency distribution of quantitative data gives the number of occurrences of the data in each interval of the data over the range of the data.
* The histogram of a given data is obtained using hist (name of the vector).
* Median is the middle most value of a set of observations, when the observations are arranged in ascending order.
* Quartile is the value of a variable of interest corresponding to a percentage of the total frequency in steps of 25 per cent of the observations.
* Percentile is the value of the variable of concern say sales with respect to a percentage from 0 per cent to a given percentage of the total frequency of the observations of that variable, when they are arranged in ascending order.
* The range of a given data specifies the difference between the maximum value of the observations and the minimum value of the observations.
* Interquartile range is the difference between the values of the variable corresponding to upper quartile and lower quartile of the observations.
* The box plot is based on quartiles and gives the lowest and highest values of the observations of a given set of observations.
* The variance of a set of observations gives the dispersion of the observations around the mean of those observations.
* The covariance of two different streams of data explains the linear dependency among them.
* The correlation coefficient is the ratio of the covariance and the product of the individual standard deviations.
* The excess kurtosis is a measure of the tail-shaped distribution.
* A probability distribution gives a curve by keeping the value of a random variable on *X*-axis and the probability of occurrence of that value of the random variable on *Y*-axis.
* The binomial distribution represents the outcome of *n* independent trials.
* The Poisson distribution represents the probability of occurrence of a specified arrival rate of customers in a queueing system (*x*).
* The exponential distribution represents the values of the probabilities for different values of the random variable *x*, where *x* may be the service time in queueing system.
* Uniform distribution represents equal probability for each value in between the lower limit and upper limit of the random variable.
* The normal distribution is a bell-shaped continuous distribution.
* The student *t* distribution is also a bell-shaped distribution like normal distribution with *n* – 1 df, where *n* is the number of observations.
* The chi-square distribution is the sum of *m* independent standard normal variables with *m* df.
* F distribution is the ratio of two chi-square distribution with n1 df and n2 df.
* Interval estimation is the process of determining the span of the random variable of a probability distribution for a given significance level.
* A hypothesis is an assumption about a population.
* The process of testing the significance of the hypothesis is called hypothesis testing.
* The chi-square test for independence is used to check whether there is dependency between levels of two categories of data.
* The goodness-of-fit test can be carried out to confirm the assumed distribution at a significance level using chi-square test.
* The completely randomized design considers only one factor with *k* treatments and *n* replications in each treatment.
* The randomized block design considers only one factor with *k* treatments and *n* replications in each treatment.
* The factorial design considers more than one factor and test the significance of each of the components of the ANOVA model at a given significance level.
* In one-tailed one-sample sign test when sample size is small, a small random sample is taken from a non-normal population and then it is tested against a median value (*μ*) such that the observations in the sample are more than that median (*μ*) or less than that median (*μ*) at a significance level of *α* using binomial distribution.
* Mann–Whitney U test is an alternate to the two samples *t* test and it is powerful, and it is based on the ranks of the combined observations of the samples.
* Kruskal–Wallis test is like the *K*-samples median test in which the objective is to test whether the *K* samples are from K identical populations.
* Simple regression aims to develop an equation, which contains an independent variable (*x*) and a dependent variable (*y*).
* The multiple regression model aims to fit a regression model to estimate the value of the dependent variable in it for a given set of independent variables, which means that the number of independent variables is more than one.

# Keywords

* **Binomial distribution** represents the outcome of *n* independent trials.
* **Box plot** is based on quartiles and gives the lowest and highest values of the observations of a given set of observations.
* **Chi-square distribution** is the sum of *m* independent standard normal variables with *m* df.
* **Completely randomized design** considers only one factor with *k* treatments and *n* replications in each treatment.
* **Correlation coefficient** is the ratio of the covariance and the product of the individual standard deviations.
* **Covariance** of two different streams of data explains the linear dependency among them.
* **Data frame** is used to store data table, which consists of a list of vectors of equal length.
* **Excess kurtosis** is a measure of the tail-shaped distribution.
* **Exponential distribution** represents the values of the probabilities for different values of the random variable *x*, where *x* may be the service time in queueing system.
* **F distribution** is the ratio of two chi-square distribution with n1 df and n2 df.
* **Factorial design** considers more than one factor and tests the significance of each of the components of the ANOVA model at a given significance level.
* **Goodness-of-fit test** can be carried out to confirm the assumed distribution at a significance level using chi-square test.
* **Interquartile range** is the difference between the values of the variable corresponding to upper quartile and lower quartile of the observations.
* **Interval estimation** is the process of determining the span of the random variable of a probability distribution for a given significance level.
* **List** contains more than one vector.
* **Matrix** is a two-dimensional array having a specified number of rows and a specified number of columns.
* **Median** is the middle most vale of a set of observations, when the observations are arranged in ascending order.
* **Multiple regression model** aims to fit a regression model to estimate the value of the dependent variable in it for a given set of independent variables, which means that the number of independent variables is more than one.
* **Normal distribution** is a bell-shaped continuous distribution.
* **Percentile** is the value of the variable of concern say sales with respect to a percentage from 0 per cent to a given percentage of the total frequency of the observations of that variable, when they are arranged in ascending order.
* **Poisson distribution** represents the probability of occurrence of a specified arrival rate of customers in a queueing system (*x*).
* **Probability distribution** gives a curve by keeping the value of a random variable on *X*-axis and the probability of occurrence of that value of the random variable on *Y*-axis.
* **Quartile** is the value of a variable of interest corresponding to a percentage of the total frequency in steps of 25 per cent of the observations.
* **R programming** is a programming software, which has the feature of having functions for most of the modules of computing, namely vector, matrix multiplication, measure of central tendencies, regression model and so on.
* **Randomized block design** considers only one factor with *k* treatments and *n* replications in each treatment.
* **Range** of a given data specifies the difference between the maximum value of the observations and the minimum value of the observations.
* **Simple regression** aims to develop an equation, which contains an independent variable (*x*) and a dependent variable (*y*).
* **Student *t* distribution** is also a bell-shaped distribution like normal distribution with *n* – 1 df, where *n* is the number of observations.
* **Uniform distribution** represents equal probability for each value in between the lower limit and upper limit of the random variable.

# Exercise

## A. Multiple Choice Questions

Refer to website

## B. Review Questions

* + - 1. Explain the steps of installing R.
      2. Illustrate adding two numbers in R.
      3. Illustrate adding two numbers after defining them through assignment statements in R.
      4. Illustrate the definition of a vector.
      5. Explain the method of handing numeric data in R using examples.
      6. Explain the method of handing integer data in R using examples.
      7. Explain the method of handling complex data in R using examples.
      8. Explain the method of handling logical data in R using examples.
      9. Explain the method of handling character data in R using examples.
      10. The members of vectors r and s are (5, 4, 5, 6, 7) and (7, 3, 4, 5, 2). Write R program to add these two vectors and print the result.
      11. What is recycling in R? Illustrate it with an example in R program.
      12. Distinguish between vector indexing and vector negative indexing through R program.
      13. What is numeric vector indexing? Illustrate it using R program.
      14. What is range indexing? Illustrate it using R program.
      15. Illustrate the method of defining a matrix and printing it in R.
      16. What is list? Illustrate member referencing in a list using R program.
      17. Consider the yield of paddy in kg in four plots using four different fertilizers as shown below.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  | Fertilizer Brand | | | |
| A | B | C | D |
| Replication  (plot) | 1 | 100 | 150 | 120 | 70 |
| 2 | 80 | 70 | 110 | 100 |
| 3 | 68 | 90 | 85 | 78 |
| 4 | 125 | 138 | 60 | 124 |

Define a data frame for the above data.

* + - 1. The hourly output of plants using different technologies are summarized below.

Hourly Output (Units)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Plant |  | Technology | | | |
|  | Technology 1 | Technology 2 | Technology 3 | Technology 4 |
| Plant 1 | 73 | 68 | 74 | 71 |
| Plant 2 | 73 | 57 | 75 | 52 |
| Plant 3 | 45 | 38 | 68 | 40 |
| Plant 4 | 73 | 41 | 75 | 75 |

* 1. Illustrate data frame of column vector.
  2. Illustrate data frame column slicing.
  3. Illustrate data frame row slicing.
     + 1. \*\* Data Import
       2. Illustrate the COUNT function for ungrouped data in R using an example.
       3. Illustrate the COUNT function for grouped data in R using an example.
       4. What is histogram? Give the syntax of constructing histogram in R.
       5. Consider the class of a professional course like MBA of a leading business school. As everybody knows that the qualifying degree for the MBA programme admission is any undergraduate degree with a specified minimum mark in that qualifying degree. Based on the admission data of the business school, it is found that the undergraduate degrees include arts, science, engineering, medicine and law. The number of students with each of these degrees in a class of 60 students is given below.

|  |  |  |
| --- | --- | --- |
| S. No. | Undergraduate Degree | Frequency of Student |
| 1 | Arts | 12 |
| 2 | Science | 9 |
| 3 | Engineering | 45 |
| 4 | Medicine | 5 |
| 5 | Law | 6 |

* + - * 1. Construct a histogram for the above data using R program.
        2. Write R program to construct cumulative frequency distribution.
        3. Write R program to construct cumulative frequency graph.
      1. The demand values in thousands of units of a product for the past 10 years are summarized in the following table. Write R program to find the mean demand of the product.

|  |  |
| --- | --- |
| Year | Demand (Thousands of Units) |
| 1 | 250 |
| 2 | 220 |
| 3 | 270 |
| 4 | 290 |
| 5 | 260 |
| 6 | 300 |
| 7 | 320 |
| 8 | 320 |
| 9 | 350 |
| 10 | 375 |

* + - 1. The sales revenue of a product during the past 10 years is summarized in the following table.

Write modules of R program to

* + - * 1. Get the range of the sales data
        2. Find the median sales of the product
        3. Find different quartile of the given sales revenue data
        4. Find 40th percentile and 74th percentile of the sales data
        5. Find interquartile range of the sales data
        6. Draw box plot for the sales data

|  |  |
| --- | --- |
| Year | Sales Revenue (` in Crore) |
| 1 | 32 |
| 2 | 36 |
| 3 | 21 |
| 4 | 52 |
| 5 | 40 |
| 6 | 32 |
| 7 | 36 |
| 8 | 40 |
| 9 | 36 |
| 10 | 38 |

* + - 1. Consider two different salesmen of a company whose annual sales figures in lakhs of rupees for the past five years are as shown below.

Write R program to

* + - * 1. Find the variance of the annual sales of each salesman
        2. Find the standard deviation of the annual sales of each salesman
        3. Find the covariance of the annual sales of the salesmen
        4. Find the correlation coefficient of the annual sales of the salesmen

|  |  |  |
| --- | --- | --- |
| Year | Salesman A  (` in Lakh) | Salesman B  (` in Lakh) |
| 1 | 85 | 100 |
| 2 | 90 | 120 |
| 3 | 95 | 50 |
| 4 | 150 | 60 |
| 5 | 200 | 125 |

27. The annual sales in crores of rupees of a company for the past 12 years are shown in the following table.

Write R program to

1. Find the Karl Pearson’s coefficient of skewness of the annual sales.
2. Find kurtosis of the annual sales data.

|  |  |
| --- | --- |
| Year | Sales in Crores of Rs. |
| 1 | 10 |
| 2 | 12 |
| 3 | 14 |
| 4 | 12 |
| 5 | 16 |
| 6 | 10 |
| 7 | 14 |
| 8 | 10 |
| 9 | 15 |
| 10 | 12 |
| 11 | 15 |
| 12 | 16 |

1. Write R program to construct the probability mass function of (a) tossing 2 coins, (b) tossing 3 coins and (c) rolling two dices.
2. Based on the past experience, the chief of a consultancy organization has estimated that the probability of completing each in time is 0.7. The company is planning to execute 12 such projects in the forthcoming quarter. Write R program to find the probability of completing
   1. No project in time
   2. Four projects in time
   3. At most 3 projects in time

* 30. The arrival rate of customers arriving at a petrol bunk follow Poisson distribution with a mean arrival rate of 10 per 15-minute interval. Write R program to find the probability that
  1. No customer will arrive in 15-minute interval
  2. Exactly 3 customers will arrive in 15-minute interval
  3. At most 3 customers will arrive in 15-minute interval
  4. At least 4 customers will arrive in 15-minute interval

31. In an international airport, the service time for serving flights by a terminal follows exponential distribution. The service rate of a terminal serving the flights is `20 per day. Write R program to find the probability that the service time of the terminal in clearing a flight is

* 1. Less than 0.45 hr
  2. More than 1 hr

32. In a survey with a sample of 350 respondents, the monthly income of the respondents follows normal distribution with its mean and standard deviation as `25,000 and `4,000, respectively. Answer the following using R program.

1. What is the probability that the monthly income is less than `20,000?
2. What is the probability that the monthly income is more than `30,000?
3. What is the probability that the monthly income is in between `22,000 and `28,000?

33. A random sample of 20 dealers of a company is taken from a normal population. The mean and variance of the annual sales of the population are `60 lakh and `110 lakh, respectively. Using R program, find the probability that the mean annual sales of the sample

* 1. Is less than `65 lakh
  2. More than `65 lakh

34. A random sample of 25 respondents is taken from a normal population. The variance of the annual income of the respondents from the normal population and that of the random sample of 25 respondents are `5 lakh and 8 lakh, respectively.

Using R, compute chi-square statistic and find the probability that the chi-square value is more than the calculated chi-square statistic.

35. Two independent samples of students of a programme under online education are taken from normal populations with the same variance. The size and variance of marks of the first sample are 12 and 114, respectively. The size and variance of marks of the second sample are 22 and 48, respectively. Using R,

* 1. Calculate *F* statistic.
  2. Probability that the F-ratio is more than the calculated *F* statistic.

37. The outer length of connecting rods produced follows normal distribution with the variance of 600 mm. Using R, determine the sample size such that the mean outer length of connecting rods is within plus or minus 1 mm with a confidence level of 0.9.

38. The monthly sales of the salesmen of Alpha company follows normal distribution. The targeted mean of the weekly sales of the salesmen is `3 lakh. The variance of the weekly sales of the salesmen is `0.8 lakh. The regional marketing manager of the company feels that the performance of salesmen has declined in the recent past. A random sample of 45 salesmen is taken from the normal population for which the weekly sales is found to be `2.5 lakh. Using R, check whether the sales revenue of the salesmen has really declined at a significance level of 0.05.

39. The weight of respondents in a survey follows normal distribution with finite population size of 1,000. The expected mean of the weight of the respondents of the population is 58 kg. The variance of the weight of the respondents of the population is 64. The researcher feels that the mean weight of the respondents is more than the expected mean of 58 kg in the recent past. A random sample of 75 respondents is taken from the normal population for which the mean weight is found to be 59 kg. Using R, check whether the mean weight has increased from the expected mean weight of 58 kg at a significance level of 0.10.

40. The monthly shopping number of respondents for a particular product in a survey follows normal distribution with finite population size of 900. The expected mean of the monthly shopping amount of the respondents of the population is `10,000. The variance of the monthly shopping amount of the respondents of the population is `350,000. The researcher feels that the mean monthly shopping amount of the respondents does not differ from the expected mean of `10,000 in the recent past. A random sample of 49 respondents is taken from the normal population for which the mean monthly shopping amount is found to be `9,800. Using R, check whether the mean shopping amount does not differ from the expected mean of `10,000 at a significance level of 0.05.

41. The weight of fertilizer bags produced in a fertilizer company follows normal distribution whose population is infinite. The expected mean of the weights of the fertilizer bags for sales of this population is 48 kg and its variance is unknown. The sales manager of the firm claims that the mean weight of the fertilizer bags is significantly more than the expected mean weight of the population. So he has selected a random sample of 49 bags and its mean and variance are found to be 47 kg and 1.5 kg, respectively. Using R, verify the intuition of the sales manager at a significance level of 0.05.

42. The weight of a drug produced in Gamma Pharmaceutical Company follows normal distribution whose population is infinite. The specified mean of the weight of the drug of this population is 75 mg and its variance is unknown.

The quality engineer of the firm claims that the mean weight of the drugs does not differ significantly from the specified mean weight of the population. So the purchase manager of Beta hospital who places order for that drug with the Gamma Pharmaceutical Company has selected a random sample of 49 drugs. The mean and variance of the sample are found to be 98 mg and 20 mg, respectively. Using R, verify the intuition of the quality manager of Gamma Pharmaceutical Company at a significance level of 0.05.

43. The quality manager of an air conditioner company has an intuition that the mean time between failures of the compressor received is at most 240 days. The quality manager wants to test his intuition on the mean time between failures of the motors. Hence, he has taken a sample of 20 compressors, whose mean time between failures and its variance is found to be 245 days and 25 days, respectively. Verify the intuition of the quality manager at a significance level of 0.1.

44. A consultant to a multi-plant organization is studying the number of years of stay of employees in two different plants (Plant X and Plant Y) for their relative difference. The variance of the number of years of stay of employees in Plant X is 25 and that of the employees in Plant Y is 36. The consultant feels that the number of years of stay of the employees in Plant X is more than that of the employees in Plant 2. To test his intuition, he has selected a sample of 64 employees from Plant X and their mean number of years of stay is found to be 12 years. Similarly, he has selected a sample of 49 employees from Plant Y and their mean number of years of stay is found to be 14 years. Using R, test the intuition of the consultant at a significance level of 0.1.

45. The weight of cement bags produced in Alpha cement company follows normal distribution. The quality assistant at the final inspection section of the company feels that the variance of the weight of the cement bags has increased from a specified maximum variance of 0.121 kg which will lead to customer complaints. Hence, he has selected a sample of 15 cement bags and found that the variance of the sample is 0.81 kg. Check the intuition of the quality assistant at a significant level of 0.05.

46. A researcher has collected the data summarizing the number of respondents in their study under each combination of the level of income and the level of qualification as shown in the following table.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Level of Qualification | | | |
|  | Under-graduation | Postgraduation | PhD |
| Level of Income | Low | 25 | 55 | 15 |
| Medium | 60 | 70 | 25 |
| High | 50 | 80 | 75 |

Using R, check whether the income is independent of the qualification while grouping the respondents at a significance level of 0.05.

47. The EPS of 150 companies in an industry follows uniform distribution. The observed frequencies of EPS values are summarized in the following table.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Observed Frequencies of Daily Demand | | | | | | | | | |
| EPS (Rs.) | 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 | 33 | 34 |
| Observed frequency | 18 | 15 | 12 | 15 | 14 | 14 | 13 | 16 | 19 | 14 |

Check whether the given data follow uniform distribution at a significance level of 0.01.

48. An experiment was run to determine whether four specific intervals of patrolling (in minutes) by a quality inspector affect the number of acceptable parts produced per hour in a machine. The corresponding data for 4 different periods with four replications are shown below.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Period of Inspection | | | |
|  | 15 min | 30 min | 45 min | 60 min |
|  | 60 | 30 | 40 | 33 |
| Replication | 25 | 60 | 22 | 10 |
|  | 40 | 26 | 30 | 48 |
|  | 24 | 38 | 39 | 35 |

Perform ANOVA as per completely randomized design and state the inference at a significance level of 5 per cent.

49. The marketing manager of a company wants to check whether the sales revenue (crores of rupees) is affected by sales region. Because there might be variability from one period to another period, they decide to use the randomized complete block design by treating periods as blocks. The corresponding data are presented in the table.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | | Sales Region | | | | | |
| A | B | C | D | E | F |
|  | 1 | 20 | 9 | 15 | 22 | 9 | 12 |
|  | 2 | 25 | 7 | 14 | 18 | 25 | 13 |
| Period | 3 | 20 | 10 | 14 | 9 | 15 | 17 |
|  | 4 | 11 | 13 | 30 | 12 | 20 | 23 |
|  | 5 | 18 | 12 | 25 | 15 | 15 | 8 |
|  | 6 | 22 | 30 | 17 | 16 | 20 | 28 |

Using R, check the following at a significance level of 0.05.

Whether there are significant differences among the sales regions in terms of sales revenue.

Whether there are significant differences among the blocks (periods) in terms of sales revenue.

50. A company wants to assess the contribution of its employees on a scale of 0–10 in terms of value addition to its business operations. So the UG qualification and work experience of the employees are considered as the factors of the experiment that is to be carried out by the company. The corresponding ratings of the employees are shown below.

|  |  |  |
| --- | --- | --- |
|  | **UG Degree** | |
| Work Experience | Engg. | Others |
| Less than 3 years | 8  7 | 4  8 |
| 3 years & above | 9  9 | 7  8 |

Using R, check the following at a significance level of 0.05.

* 1. Whether there are significant differences among the degrees in terms of rating of contribution by employees.
  2. Whether there are significant differences among work experience in terms of rating of contribution by employees.
  3. Whether there are significant differences among the interaction terms of degree and work experience in terms of rating of contribution by employees.

51. The age (in years) of 9 randomly selected respondents in a survey are 52, 28, 70, 65, 25, 46, 65, 42 and 52. Check whether the age of respondents is 46 years (H0:u = 46) as against the alternate hypothesis H1:μ < 46 using sign test at a significance level of 0.01.

52. The quarterly sales revenues (in lakhs of rupees) of randomly selected samples of salesmen from each of two regions are summarized below.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Region 1 | 57 | 46 | 27 | 20 | 13 | 43 | 10 | 60 | 45 | 55 | 61 | 38 | 43 | 49 | 29 | 20 | 15 | 52 | 43 | 36 |
| Region 2 | 48 | 25 | 62 | 41 | 25 | 32 | 15 | 47 | 34 | 20 | 53 | 67 | 23 | 59 | 44 | 20 | 24 | 32 | – | – |

Check whether the two samples are drawn from identical populations against the alternate hypothesis that the first population is stochastically larger than the second population using Mann–Whitney U test at a significance level of 0.05.

53. The sales revenues of a product made by three salesmen are shown in the following table.

-------------------------------------------------------

Sales revenue (` in Crore)

-------------------------------------------------------

Salesman 1 Salesman 2 Salesman 3

-------------------------------------------------------

74 95 79

82 83 86

69 72 63

71 87 72

82 91 90

90 76 89

73 79 83

85 58 68

Using R, check whether there are significant differences among the salesmen in terms of sales revenue using Kruskal–Wallis Test at a significance level of 0.05.

54. The annual training expenditure (lakhs of rupees) and the corresponding sales revenues for the past 8 years of a company are presented below.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Year (i) |  | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| Annual training expenditure, Xi | 5 | 7 | 9 | 10 | 12 | 15 | 18 | 20 |  |
| Sales revenue (` crore) Yi | 80 | 90 | 75 | 85 | 95 | 70 | 95 | 60 |  |

Using R, find the correlation coefficient between annual training expenditure and sales revenue.

55. The demand values (in lakhs of units) of a product during the past 9 years are summarized below. Using R, fit a linear regression to estimate the demand of the product.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Year (X) | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| Demand (Y) | 22 | 27 | 38 | 42 | 55 | 72 | 78 | 90 | 105 |

56. The annual sales revenue (in crores of rupees) of a product as a function of time period (year) and annual advertising expenditure (in lakhs of rupees) for the past 5 years are summarized below.

|  |  |  |
| --- | --- | --- |
| Annual Sales Revenue (Y) | Year (X1) | Annual Advertising Expenditures (X2) |
| 100 | 1 | 21 |
| 110 | 2 | 25 |
| 125 | 3 | 30 |
| 135 | 4 | 35 |
| 145 | 5 | 40 |

Using R, design a multiple regression model to forecast the annual sales revenue of the product.

# References

Panneerselvam, R. 2012. *Design and analysis of experiments*. New Delhi: PHI Learning Private Limited.

———. 2014. *Research Methodology*, 2nd ed. New Delhi: PHI Learning Private Limited.

R Tutorial. n.d. *R Introduction*. Available at <http://www.r-tutor.com/r-introduction> (accessed on 22 November 2020).

Venables, W. N., D. M. Smith, and the R Core Team. *An Introduction to R*. Available at <https://cran.r-project.org/doc/manuals/r-release/R-intro.pdf> (accessed on 22 November 2020).