ISSN 2581 – 3781

AJCSE
ASIAN JOURNAL OF COMPUTER SCIENCE ENGINEERING

Available Online at www.ajcse.info
Asian Journal of Computer Science Engineering 2018;3(4):22-26

RESEARCH ARTICLE

Statistical Analysis and Data Analysis using R Programming Language: Efficient and Flexible Evaluation

B. Usharani

Department of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, Andhra Pradesh, India

Received on: 01-09-2018; Revised on: 01-10-2018; Accepted on: 25-10-2018

ABSTRACT

R is an integrated suite of software facilities for data manipulation, data visualization, and graphical facilities. R has an effective data handling and storage facility. R provides a large, coherent, integrated collection of intermediate tools for data analysis. R provides a rich graphical facility for data analysis. R behaves like a vehicle for newly developing methods of interactive data analysis. R can use as a statistics system. R will give minimal output and store the results in a fixed object. R is becoming the leading language in statistics. R is designed to make data analysis and statistics easier to do. R is not only entrusted by academic but also many companies also use R programming language including Google, Facebook, Uber, and so on.

Key words: Data analysis, data manipulation, data visualization, graphics, statistical analysis

INTRODUCTION

R is a programming language mainly used for scientific research, data analytics, and statistical computing [Figure 1]. R is one of the programming languages used by the statisticians, data analyst, researchers and marketers toretrieve, clean analyses, visualize, and present data. Nowadays, R is the first choice of statisticians and mathematicians, professional programmers prefers implementing a new algorithm in a programming language. The main advantage of the R is getting things done with a very little code. R runs on all platforms. R programs compiles runs on Unix platforms and other systems including Linux, Windows, and MAC OS.

R is a type of software facility used for data manipulation, calculation, and graphical display. It includes a variety of uses to handle data.

R offers an effective data handling and storage facility, a suite of operators for calculations on arrays, matrices a large integrated collection of intermediate tools for data analysis, graphical facilities for data analysis, and display either on screen or on hardcopy and a well-developed simple and effective programming language which include conditions, loops, user defined recursive functions, and input and output facilities.

The best algorithms for machine learning can be implemented with R.R can communicate with the other language. R can call python, java, and Cpp code.

The R can be dived into four parts called analytics, graphics, application, and programming language. The analytics is subcategorized as statistics, probability distributions, Big data analytics, machine learning, optimization and mathematical problems, signal processing, statistical modeling and statistical tests, and simulation random number generation. The graphics is subcategorized as static graphics, dynamic graphics, and interactive graphics. The application is subcategorized as applications, data mining, and statistical methodology. The R programming language is object oriented, procedural, scripting, and interpreted language.

The R system can be divided into two parts. One is the base R system that can be downloaded from CRAN [Figure 2]. The base R contains the bae packages which are required to run R and contains the most fundamental functions and include packages such as utils, stats, datasets, graphics, grid, tools, parallel, compiler, splines, stringr, class, and cluster. There are about 7800 packages on CRAN that have been developed by users and programmers around the world.

Address of correspondence:
B. Usharani
E-mail: usharedddy.vja@gmail.com

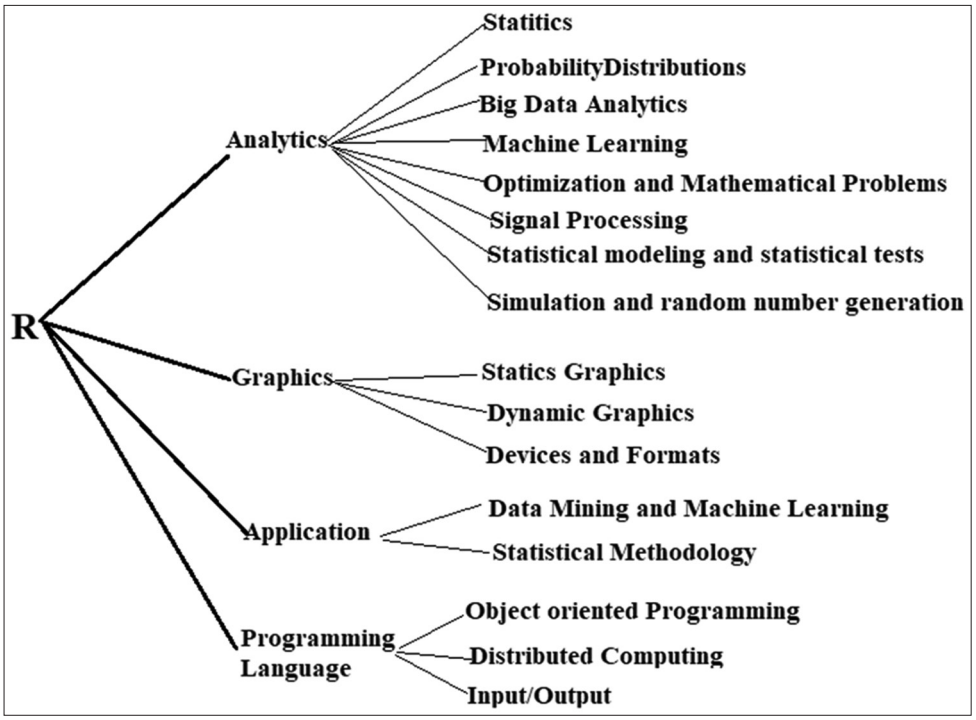


Figure 1: R programming language summarization

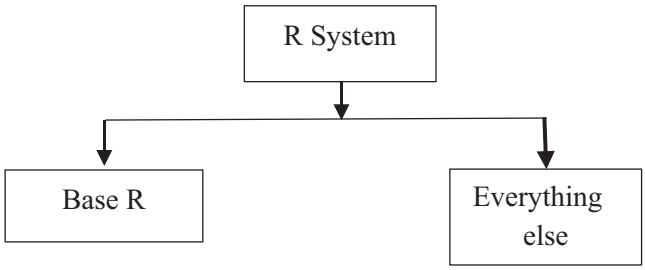


Figure 2: R system

R history

The "R" name is derived from the first letter of the names of its developers Ross Ihaka and Robert Gentleman who were associated with the University of Auckland at the time.[1] The initial version of R was released in 1995. R is an implementation of the "S" programming language[2] and combines with lexical scoping semantics. R is best for statistical, data analysis and machine learning. R software environment was written in C, Fortran and R. 50% of R is written in C and 30% in Fortran.[3,4]

Advantages of R

1. R is an open source software.
2. R is platform independent.
3. R includes many packages to provide technologies.
4. Allows users to add additional functionality by defining new functions.
5. C, C++, and some other programming code can be linked and called at runtime.
6. R is the only statistical open source software.
7. R has an effective data handling and storage facilities.

Table 1: Linear regression program

Program	Output
x <- rnorm (50, mean = c (rep (0, 25), rep (3, 25)))	none Bonferroni FDR BH
p <- 2 * pnorm (sort(-abs (x)))	[1,] 0.000 0.000 0.000 0.000
pVal<- round (p, 3)	[2,] 0.000 0.001 0.000 0.000
Bonferroni <- round (p.adjust (p, "bonferroni"), 3)	[3,] 0.000 0.002 0.001 0.001
## FDR and BH are equivalent	[4,] 0.000 0.003 0.001 0.001
FDR <- round (p.adjust (p, "fdr"), 3)	[5,] 0.000 0.007 0.001 0.001
BH <- round (p.adjust (p, "BH"), 3)	[6,] 0.000 0.010 0.002 0.002
res <- cbind (none = pVal,	[7,] 0.000 0.012 0.002 0.002
Bonferroni = Bonferroni, FDR = FDR,	[8,] 0.000 0.019 0.002 0.002
BH = BH)	[9,] 0.001 0.027 0.003 0.003
res <- res[order (res[, "Bonferroni"]),]	
print (res[1:10,])	[10,] 0.001 0.029 0.003 0.003

Table 2: Calculating statistical parameters program

Program	Output
Mode <- function (x) {	[1] "Mean : 8.220000"
ux<- unique (x)	[1] "Median: 5.600000"
ux[which.max (tabulate (match (x, ux)))]	[1] "standard deviation: 19.200567"
}	[1] "Varaince: 368.661778"
x <- c (12,7,3,4.2,18,2,54,-21,8,-5)	0% 25% 50% 75% 100%
sprintf("Mean : %f", mean (x))	-21.00 2.25 5.60 11.00 54.00
sprintf("Median:%f," median (x))	[1] "Mode: 12.000000"
sprintf("standard deviation:%f," sd (x))	
sprintf("Varaince:%f," var (x))	
print (quantile (x))	
sprintf("Mode:%f," Mode (x))	

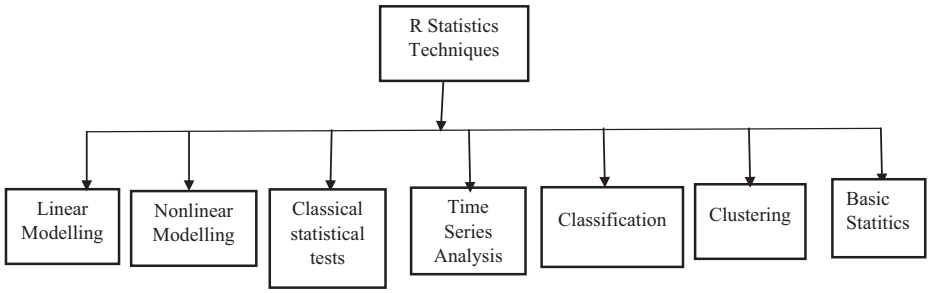


Figure 3: R statistics techniques

Table 3: Plotting the data example in R

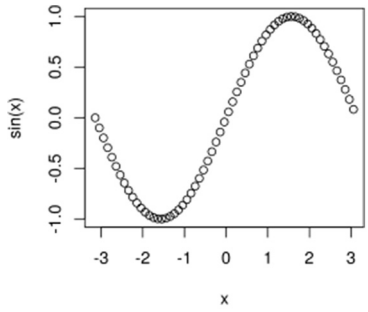
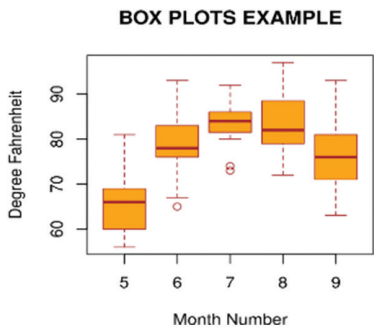
Program	Output
x <- seq(-pi, pi, 0.1) plot (x, sin (x))	

Table 4: Box plots example in R

Program	Output
Boxplot (Temp ~ Month, Data = Airquality, Main="BOX PLOTS EXAMPLE," Xlab="Month Number," Ylab="Degree Fahrenheit," Col="orange," Border="Brown"	

8. R has facilities to print the reports for the analysis performed in the form of graphics.
9. R has the capabilities for parallel computation.

Disadvantages of R

1. R can consume all available memory.
2. In R, no one to complain if something does not work.
3. R is very slow
4. In R quality of some packages is lessperfect.
5. Variables are strictly local in scope.

R statistics techniques

Statistics includes methods of collecting, organizing, and analyzing data in such a way that meaningful conclusions can be drawn from them [Figure 3]. R is a robust environment for analysing the data.[5] Statistical learning emphasizes models and their interpretability and precision and uncertainty. In statistics, nonlinear is a form of regression analysis in which observational data are modeled by a function which is a nonlinear combination of the model parameters and depends on one or more independent variables. Classification is a technique that assigns categories to a collection data to do prediction and analysis. The in-built methods to do statistical computing are mean, mode, median, quartile, variance, standard deviation, cross tabulation, and correlation [Figure 4].

Table 5: Data summary example in R

Program	Output
A <- data.frame (a = LETTERS[1:10], x = 1:10) class(A) sapply(A, class) typeof(A) names(A) dim(A) head(A) tail(A, 1) head(1:10, −1)	[1] "data.frame" a x "factor" "integer" [1] "list" [1] "a" "x" [1] 10 2 a x 1 A 1 2 B 2 3 C 3 4 D 4 5 E 5 6 F 6 a x 10 J 10 [1] 1 2 3 4 5 6 7 8 9

Table 6: Data summary example in R

Program	Output
a <- rnorm (50)	[1] -2.466525
min (a);	[1] 1.501929
max (a)	[1] -2.466525 1.501929
range (a)	Min. 1st Qu. Median Mean 3rd Qu. Max.
summary (a)	-2.46652 -0.76518 -0.02978 -0.08924 0.77485 1.50193

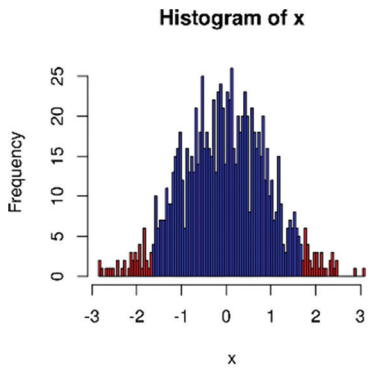
Statistical modeling in R example

The statistical modeling is that R is explained with the help of some program examples [Tables 1-8].

Table 7: Data summary in-built methods example in R

Program	Output
m <- matrix (c(1:10, 11:20), nrow = 10, ncol = 2) apply (m, 1, mean) l <- list (a = 1:10, b = 11:20) lapply (l, mean) l.mean<- sapply (l, mean) class (l.mean) attach (iris) tapply (iris$Petal.Length, Species, mean) by (iris[, 1:4], Species, colMeans)	1] 6 7 8 9 10 11 12 13 14 15 $a [1] 5.5 $b [1] 15.5 [1] "numeric" setosa versicolor virginica 1.462 4.260 5.552 Species: setosa Sepal.LengthSepal.WidthPetal.LengthPetal.Width 5.006 3.428 1.462 0.246 Species: Versicolor Sepal.LengthSepal.WidthPetal.LengthPetal.Width 5.936 2.770 4.260 1.326 Species: Virginica Sepal.LengthSepal.WidthPetal.LengthPetal.Width 6.588 2.974 5.552 2.026

Table 8: Plotting the histograms using R

Program	Output
x <- rnorm (1000) hx <- hist (x, breaks = 100, plot = FALSE) plot (hx, col = ifelse (abs (hx$breaks) < 1.669, 4, 2))	

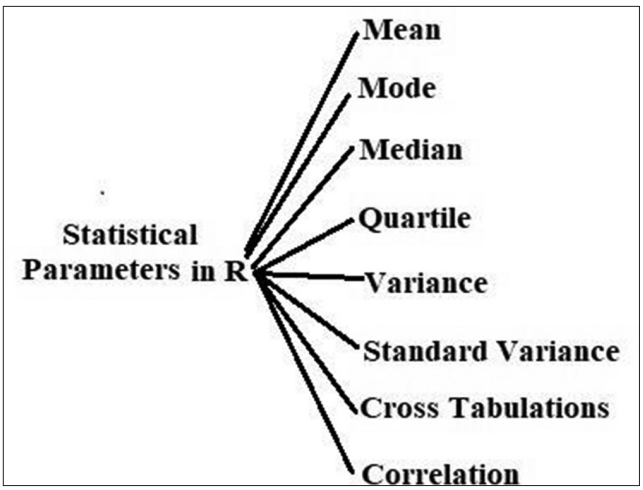


Figure 4: Statistical parameters in R

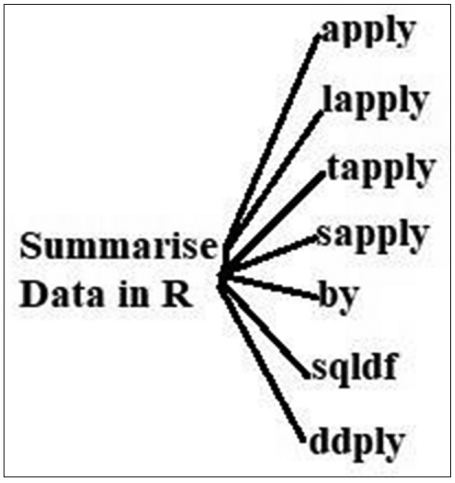


Figure 5: Summarization methods in R

Data analysis

Data analysis is an approach for summarizing and visualizing the important characteristics of a data set [Figure 5].

Advantages of data analysis using R

1. Simple and advanced options of analysis available.
2. Easy to learn.

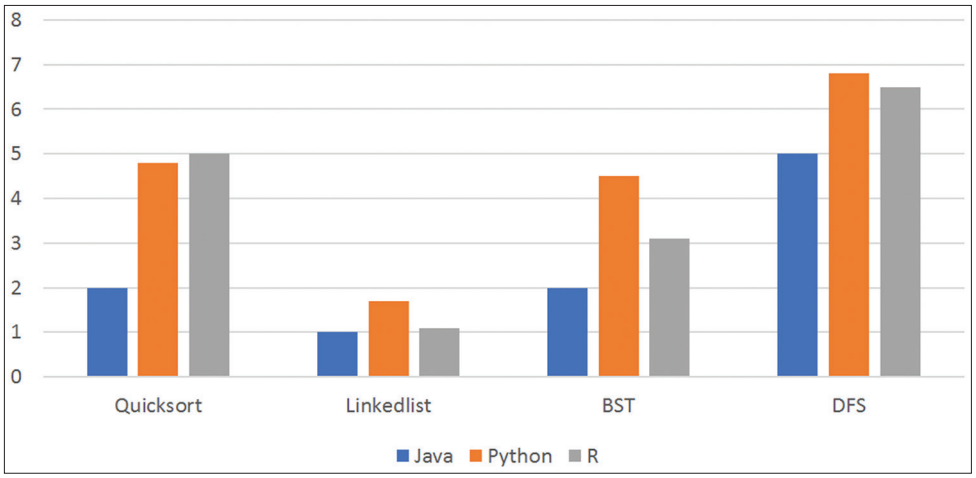


Figure 6: Evaluating the execution times in Java, Python, and R

3. Straight forward handling of analyses using simple calculations.
4. Flexible.
5. Provides both application and statistical area.
6. Ability to easily fix bugs.
7. Increases efficiency.

Data visualization

R has inbuilt plotting commands to create graphs.

Data summarization

The summarize in-built methods are apply, lapply, tapply, sapply, by, sqldf, and ddply. An example is given for these methods.

Evaluating execution time of R, java, and python

It is calculated the execution times for sorting, linked list, bst, and dfs programs in Java, Python, and R. The execution times are plotted in the form of the chart as shown in Figure 6:

When compared to java, R is slow. However, sometimes R is almost equal to the execution time of the python language.

CONCLUSION

R is a well-developed, simple, and effective programming language. Many users will come to R due to its graphical facilities. R becomes the leading language in data science. R is the tool of choice for data science professionals, especially in the IT industry. Many data analysts and research programmers use R because R is the most prevalent programming language. R is also use as a fundamental tool for finance. Companies hiring R, IBM, Aricent, KPIT, BOSCH, and Global Logic. Industries that are using R, Social Media, Public Affairs, Services, Analytics, Software Vendor Revolution Analytics, Finance, and Media are popular in application areas of Bioinformatics.

REFERENCES

1. Available from: https://www.en.wikipedia.org/wiki/R_(programming_language). [Last accessed on 2018 Sep 14].
2. Ihaka R, Gentleman R. R: A language for data analysis and graphics. J Comput Graph Stat 1996;5:299-314.
3. Available from: https://www.blog.revolutionanalytics.com/2011/08/what-language-is-r-written-in.html. [Last accessed on 2018 Sep 14].
4. Wrathematics How Much of R is Written in R. Available from: http://www.librestats.com/2011/08/27/how-much-of-r-is-written-in-r/. [Last accessed on 2018 Sep 14].
5. Thieme N. R generation. Significance 2018;15:14-9.