Syllabus

Unit - I : *Learn how to load data, plot a graph viz. histograms (equal class intervals and unequal class intervals), box plot, stem-leaf, frequency polygon, pie chart, ogives with graphical summaries of data.*

Unit - II : *Generate automated reports giving detailed descriptive statistics, correlation and lines of regression.*

Unit - III : *Random number generation and sampling procedures. Fitting of polynomials and exponential curves. Application Problems based on fitting of suitable distribution, Normal probability plot.*

Unit - IV : *Simple analysis and create and manage statistical analysis projects, import data, code editing, Basics of statistical inference in order to understand hypothesis testing and compute p-values and confidence intervals.*

Chapter 1

Various Statistical Graph Plotting

1.1 Using R

▶ 'R' Programming Language is a powerful tool to perform various statistical techniques.
▶ Its easy to learn and we can perform several statistical ideas in a very short time.
▶ We will study step by step to learn R programming.
▶ You need a desktop/ laptop/ mobile (in worst case) to perform the computation.

1.1.1 Installation

■ If you have a desktop/ laptop then perform the following steps to install 'R'.
▶ Step 1: Type **'download R programming language'** in Google. The first link will look like: **'https://cran.r-project.org/bin/windows/base/'**.
▶ Step 2: Click the download option and download the '.exe' file. For me it was **'Download R 4.1.1 for Windows'** with size around 85 mb.
▶ Step 3: Download **'RStudio Desktop'** (Version RStudio-1.4.1717, size around 156 mb) using Google or, from the link **'https://www.rstudio.com/products/rstudio/download/**.
▶ Step 4: Install 'R' and then 'R Studio' in your device.
▶ Step 5: Find 'R Studio' icon and open.
▶ Step 6: Initially you will found 3 panels: left is console, upper right is environment and lower right is plot section (default).
▶ Step 7: In console just write: '2+3' and push enter. If you get 5 then you may assume that all are going right.

■ If you have a mobile (android) then perform the following steps to install 'R'.
▶ Step 1: Type **'R Programming Compiler'** in Playstore and Install.
▶ Step 2: Delete sample program and write: '2+3' and tap green triangle. If you get 5 then you may assume that all are going right.

1.1.2 R Script

■ PC: In top left: '$File -> New\ File -> R\ Script$'. We will use 'R Script' files to write and store our code. Note that now 4 panels appear in R Studio and it is the standard format. Press 'Ctrl + s' and give name say 'Code1' and save the R script. We will write and store our code in that file.

■ Mobile: In top right click 'three dots': 'New' then name it say 'Code1' and save. A new 'R Script' file will open 'Code1.r' and we will write and store our code in that file.

1.1.3 Basics

- *Package Install & Clear Editor:* $library()$, $install.packages("MPV")$.[If does not work then install it manually]. In RStudio go to $Tools -> Install\ Packages -> (write\ package\ name\ in\ package\ archieve)$. To clear console: $Ctrl + L$. To clear all variables: $rm(list = ls())$, To remove a: $rm(a)$.
- *Easy Mathematical Operators: Addition* $[2 + 3]$, *Power* $[2 \wedge 4]$, $log(3)$, $log10(3)$, $sin(pi/2)$, $cos(pi)$, $tan(0)$, $exp(2)$, $factorial(5)$, $choose(4, 2)$, $beta(2, 3)$, $gamma(4)$.
- *Define and Use Variable:* $x = 2, y = 3$, *Addition* $[x+3*y]$, *Power* $[x \wedge y]$, $log(x*y)$, $log10(x)$, $factorial(x+y)$, $choose(y, x)$, $beta(2 + x, 3 - x)$, $gamma(y)$.
- *Numeric Vector:* $a = c(1, 3, 5, 2)$, *Addition* $[a + 2]$, *Power* $[a \wedge 2]$, $log(a)$, $factorial(a)$, $beta(a, a + 1)$, $gamma(a)$, $length(a)$, $a[3]$, $a[c(1, 4)]$.
- Character Vector: $a = c(1, 3, 5, 2)$, $a > 2$, $a > 2\ \&\ a < 4$, $a[a > 2]$, $a = c(1, 2, 3, 4)$, $b = c(2, 4, 6, 8)$, $b[a > 3]$, $b = c("a1", "a2", "a3")$, $cat(b, "\n")$, $b = c("a1", "a2", 2)$, $b[3]$, $b[3] + 2$, $as.numeric(b[3]) + 2$.
- Create Vector: *A)* $a = c(1, 2, 3)$, $b = c(3, 4, 5)$, $c = c(a, b, 2, 3)$. *B)* $a = seq(1, 10)$, $a = seq(1, 10, 2)$, $a = seq(1, 10, 0.2)$, $a = 1 : 10$. *C)* $rep(a, 2)$, $rep(a, a)$.
- Functions and Plot: $a = c(1, 3, 5, 2)$, $mean(a)$, $var(a)$, $sd(a)$, $sum(a)$, $sort(a)$, $order(a)$, $plot(a)$, $b = c(1.2, -1.3, 2.6, 7.2)$, $plot(a, b)$, $lines(a, b)$.
- Matrix: $a = matrix(1 : 20, nrow = 4)$, $dim(a)$, $t(a)$, $rownames(a) = letters[1 : 4]$, $colnames(a) = LETTERS[1 : 5]$, $a = matrix(1 : 20, ncol = 4)$, $a = matrix(1 : 20, nrow = 4, byrow = T)$, $A = cbind(a = 1 : 3, b = 4 : 6, c = 7 : 9)$, $B = rbind(a = 1 : 3, b = 4 : 6, c = 7 : 9)$, $Multiply :\ A\%*\%B$, $Multiply\ Elementwise :\ A * B$, $Inverse :\ solve(A)$, $Determinant :\ det(A)$.

1.1.4 Data Loading

■ Type 1:

► a) Small Numeric Vector e.g. $1, 2, 5, 3, 6, 9, 7$ can be easily loaded into R by typing: $a = c(1, 2, 5, 3, 6, 9, 7)$.

b) Large Numeric Vector from a '.txt' or, '.xl' or, '.csv' can be easily loaded into RStudio: $File -> Import\ Dataset$.

Example 1.1. *Create a '.txt' file contains the data: 64, 78 48 11 47 50 47 06 63 34 22 43 77 76 66 39 44 34 84 85 24 66 18 20 10 45 62 96 09 44. Import it in R and store in a variable say 'y'.*

► *R Code: Create the '.txt' file (one column vector data say 'Stat.txt' with first row name x and other rows are the data). Now go to '$File -> Import\ Dataset -> From\ Text -> Locate\ Stat.txt -> Import$. It will open the data (close it). Copy the code shown in the console $[Stat <- read.csv("C : /Users/\cdots/Stat.txt", sep = " ")]$ into your R Script and type 'Stat' but you will see this as a 'Data Frame'. To handle its element usually we use two methods: i) Attach-detach ii) Use 'Stat$x'.*

Write '$y = Stat\$x,\ y$' in R Script and Run.

► *You can write the data in excel file and convert it (save as $->$ comma delimited) into a '.csv' file. The import process is same as above.*

Example 1.2. *Create a '.xlsx' file (excel file) contains the data: 64, 78 48 11 47 50 47 06 63 34 22 43 77 76 66 39 44 34 84 85 24 66 18 20 10 45 62 96 09 44. Import it in R and store in a variable say 'y'.*

► *R Code: Create the '.xlsx' file (one column vector data say 'Stat.xlsx' with first row name x and other rows are the data). Now go to 'File − > Import Dataset − > From Excel − > Locate Stat.xlsx − > Import. It will open the data (close it). Copy the code shown in the console (except any warning massage and last line) into your R Script and type 'Stat' but you will see this as a 'Data Frame'. To handle its element usually we use two methods: i) Attach-detach ii) Use 'Stat$x'.*
Write 'y = Stat$x, y' in R Script and Run.

[Do It Yourself] 1.1. *Create a '.xlsx' file (excel file) contains the data: 64, 78 48 11 47 50 47 06 63 34 22 43 77 76 66 39 44 34 84 85 24 66 18 20 10 45 62 96 09 44 in first column and you write a second column of your own with the same length. Import it in R and store in a two variable say 'y_1, y_2'.*
[Hint : Suppose name of the Data Frame is A with column names y_1, y_2. Now 'attach(a), y_1, y_2, detach(a)'].
► *Note that, attach(a) must end with detach(a) if you write attach(a) multiple times then there will be some problems, so be careful about it.*

[Do It Yourself] 1.2. *Write down a matrix in R of 10 rows and 4 columns with your own data. How do you access of its particular rows and columns?*

[Do It Yourself] 1.3. *Write down a matrix in excel of 10 rows and 4 columns with 4 column names. Now import this matrix into R as a data frame. How do you access its particular column?*

■ Type 2:
► a) Suppose you have two or, more column (row) vectors then use 'cbind(rbind)' to join them.

Example 1.3. *Create three column vectors,* $a_1 : 64, 78, 48, 11, 47, 50, 47, 06, 63, 34$; $a_2 : 22, 43, 77, 76, 66, 39, 44, 34, 84, 85$; $a_3 : 24, 66, 18, 20, 10, 45, 62, 96, 09, 44$. *Create a new data joining this three columns.*

► *R Code:*

```
a1=c(64,78,48,11,47,50,47,06,63,34)
a2=c(22,43,77,76,66,39,44,34,84,85)
a3=c(24,66,18,20,10,45,62,96,09,44)
a=cbind(a1,a2,a3)
summary(a)
boxplot(a)
```

■ Type 3: Categorical Data (Loading and Representation)
▶ Categorical data are such that measurement scale consists of a set of categories.
▶ Marital status: never married, married, divorced, widowed (nominal or, no order).
▶ Hair Color: black, white, golden, red (nominal or, no order).
▶ Economic Status: poor, middle, rich (ordinal or, there is some order).
▶ Grade of a Student: bad, average, good (ordinal or, there is some order).

Example 1.4. *Suppose you have a categorical data with* 1 *variable in Table 1.1. Represent the data graphically.*

Categorical Data (1 Variable)					
	Never Married	Married	Divorced	Widowed	Total
Marital Status	180	210	70	40	500

Table 1.1: Marital Status Data.

▶ *R Code*:

```
status=matrix(c(180,210,70,40),nrow=1,ncol=4,byrow=T)
rownames(status)=c("Marital Status")
colnames(status)=c("Never married","Married","Divorced","Widowed")
status
barplot(status,beside=T,main="Graphical Representation",
legend.text=rownames(status),col='green')
```

▶ *R Plot*: *See the plot Fig. 1.1.*

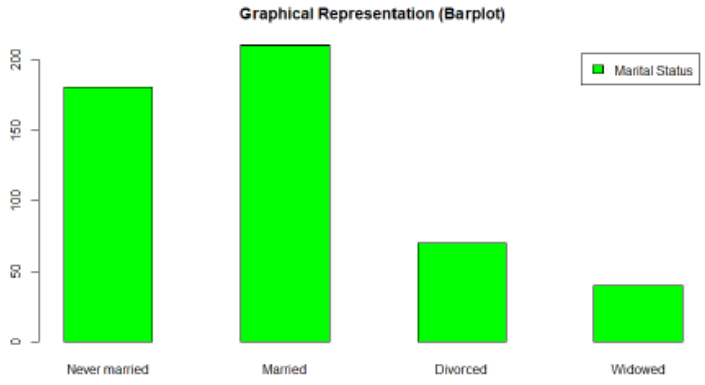

Figure 1.1: Boxplot for the data of Table 1.1.

Example 1.5. *Suppose you have a categorical data with* 2 *variables in Table 1.2. Represent the data graphically.*

Categorical Data (2 Variables)					
$\frac{MaritalStatus}{Income}$	Never Married	Married	Divorced	Widowed	Total
Low	180	210	70	40	↓
High	120	330	140	60	1150

Table 1.2: Marital Status Data.

▶ *R Code*:

```
status=matrix(c(180,210,70,40,120,330,140,60),nrow=2,ncol=4,byrow=T)
rownames(status)=c("Marital Status","Income")
colnames(status)=c("Never married","Married","Divorced","Widowed")
status
barplot(status,beside=T,main="Graphical Representation",
legend.text=c("low","High"),col=c('green','red'))
```

▶ *R Plot*: *See the plot Fig. 1.2.*

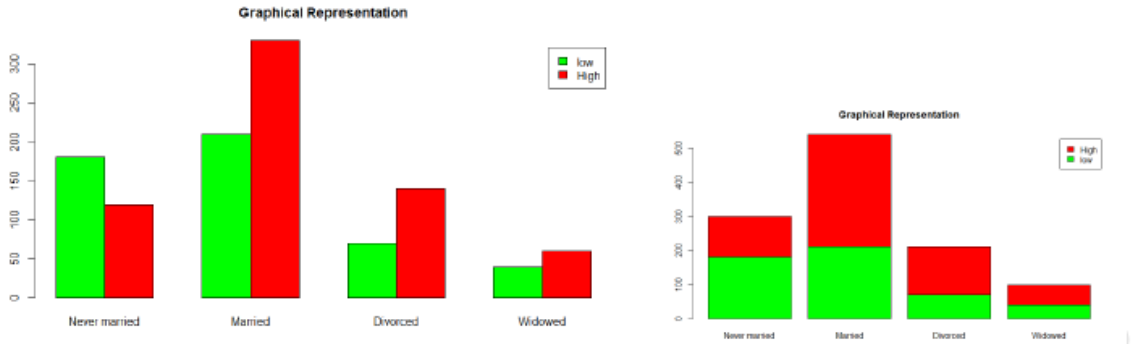

Figure 1.2: Multiple Barplot for the data of Table 1.2. a) beside=T b) beside=F

Example 1.6. *For the categorical data with* 2 *variables in Table 1.2. Represent the data graphically by Mosaic Plot.*

▶ *R Code*: *The data in Table 1.2 can be represented in excel as given in Fig. 1.3 lower panel. Now import the excel file in a data frame 'U' and use the code below.*

```
U
U1 = xtabs(Freq~MarStat+Income, data=U)
U1
mosaicplot(~MarStat+Income, data= U1)
mosaicplot(~MarStat+Income, col = c("firebrick", "goldenrod1"), data= U1)
```

▶ *R Plot*: *See the plot Fig. 1.4 (upper left).*

Sex	Hair	Eye Colour: Brown	Blue	Hazel	Green
M	Black	32	11	10	3
	Brown	53	50	25	15
	Red	10	10	7	7
	Blond	3	30	5	8
F	Black	36	9	5	2
	Brown	66	34	29	14
	Red	16	7	7	7
	Blond	4	64	5	8

	A	B	C
1	MarStat	Income	Freq
2	Never Married	Low	180
3	Married	Low	210
4	Divorced	Low	70
5	Widowed	Low	40
6	Never Married	High	120
7	Married	High	330
8	Divorced	High	140
9	Widowed	High	60

	A	B	C	D
1	Hair	Eye	Sex	Freq
2	Black	Brown	Male	32
3	Brown	Brown	Male	53
4	Red	Brown	Male	10
5	Blond	Brown	Male	3
6	Black	Blue	Male	11
7	Brown	Blue	Male	50
8	Red	Blue	Male	10
9	Blond	Blue	Male	30
10	Black	Hazel	Male	10
11	Brown	Hazel	Male	25
12	Red	Hazel	Male	7
13	Blond	Hazel	Male	5
14	Black	Green	Male	3
15	Brown	Green	Male	15
16	Red	Green	Male	7
17	Blond	Green	Male	8
18	Black	Brown	Female	36
19	Brown	Brown	Female	66
20	Red	Brown	Female	16
21	Blond	Brown	Female	4
22	Black	Blue	Female	9
23	Brown	Blue	Female	34
24	Red	Blue	Female	7
25	Blond	Blue	Female	64
26	Black	Hazel	Female	5
27	Brown	Hazel	Female	29
28	Red	Hazel	Female	7
29	Blond	Hazel	Female	5
30	Black	Green	Female	2
31	Brown	Green	Female	14
32	Red	Green	Female	7
33	Blond	Green	Female	8

Figure 1.3: Multiple Barplot for the data of Table 1.2. a) beside=T b) beside=F

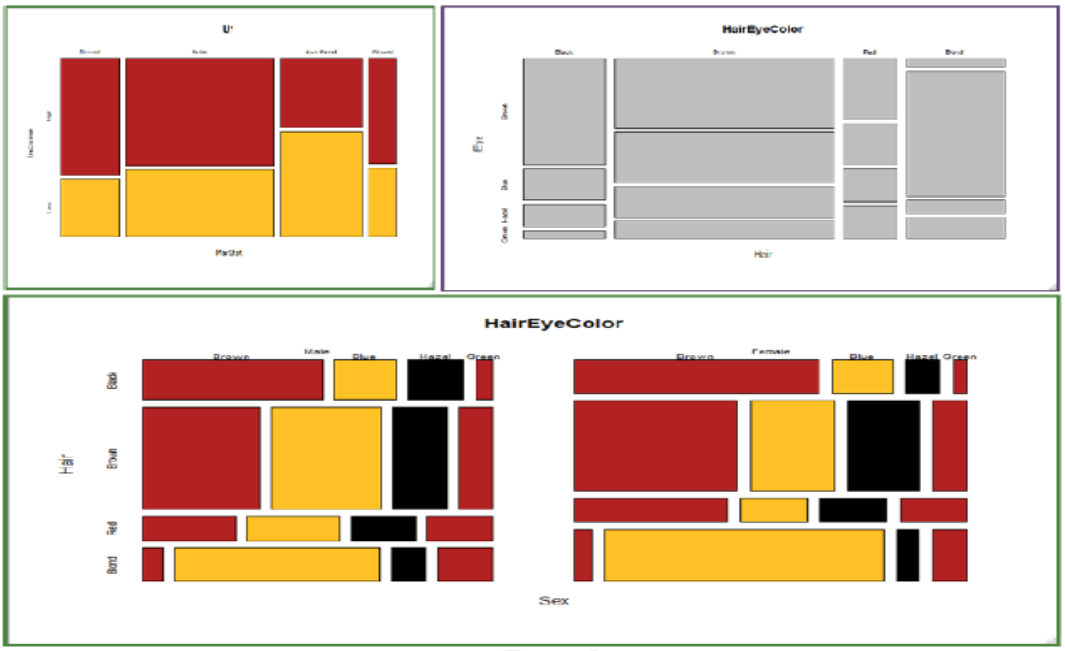

Figure 1.4: Multiple Barplot for the data of Table 1.2. a) beside=T b) beside=F

■ Type 4: High Dim. Categorical Data (Loading and Representation [Mosaic Plot])

Example 1.7. *For the categorical data with* 3 *variables in Fig. 1.3. Represent the data graphically by Mosaic Plot.*

▶ *R Code: The data in Fig. 1.3 (left) can be represented in excel as given in Fig. 1.3 right panel. Now import the excel file in a data frame 'U' and use the code below.*

```
U
U1 = xtabs(Freq~Hair+Eye+Sex, data=U)
U1
mosaicplot(~Hair+Eye, data= U1)
mosaicplot(~ Sex+Hair+Eye, col = c("firebrick", "goldenrod1", "black"), data= U1)
```

▶ *R Plot: See the plot Fig. 1.4 upper right and below.*