R-Programming

R Programming Workshop (PART 1)

Prepared by Chan Weng Howe

Outline

- Introduction to R
- R Basics
- Packages in R
- Search Help in R
- Functions in R
- Control Structures in R
- Import/Export in R
- Graphics with R

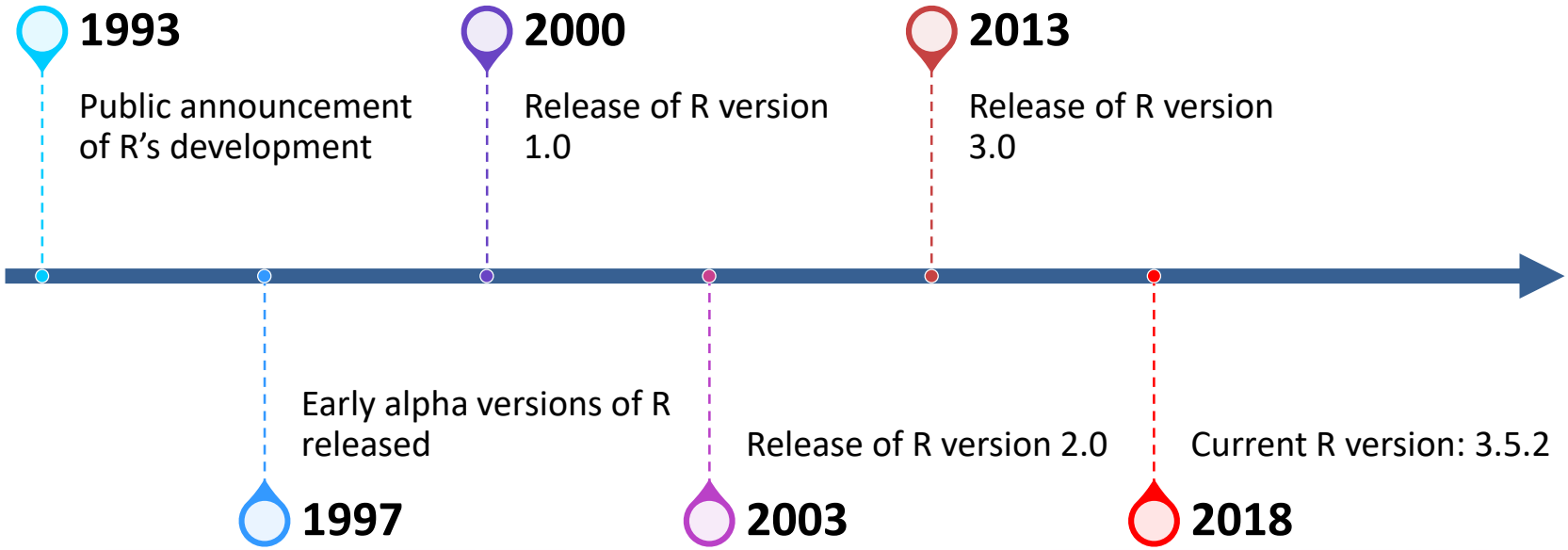
INTRODUCTION TO R

R……a mythical beast?

- R is **free** and **open-source**
- R is a **command-line** programming language
- It is not a statistics program! But a programming language that **works well for statistics**
- With **visualization capabilities** and **highly extensible**.
- Operating system independent (supported in: Windows / macOS and Linux)
- R can be integrated with other packages and languages (Excel, SAS, SPSS, etc.)

A Brief History

- R was developed by **Ross Ihaka** and **Robert Gentleman**
- The original goal was to develop software for students to use
- They were encouraged to make it an open source project.



Using R

- 2 components:
  - R Programming environment (https://www.r-project.org/)
  - RStudio IDE for R (https://www.rstudio.com/products/rstudio)
- R Programming environment provides the **core files** and **libraries** of R
- RStudio provides an **additional layer of GUI** which is consistent across Windows, macOS and Linux.

Installation of R 3.5.2

- R 3.5.2 can be downloaded at:
  - https://cloud.r-project.org/bin/macosx/R-3.5.2.pkg (macOS)
  - https://cloud.r-project.org/bin/windows/base/R-3.5.2-win.exe (Windows)
- This should be installed before the installation of RStudio.

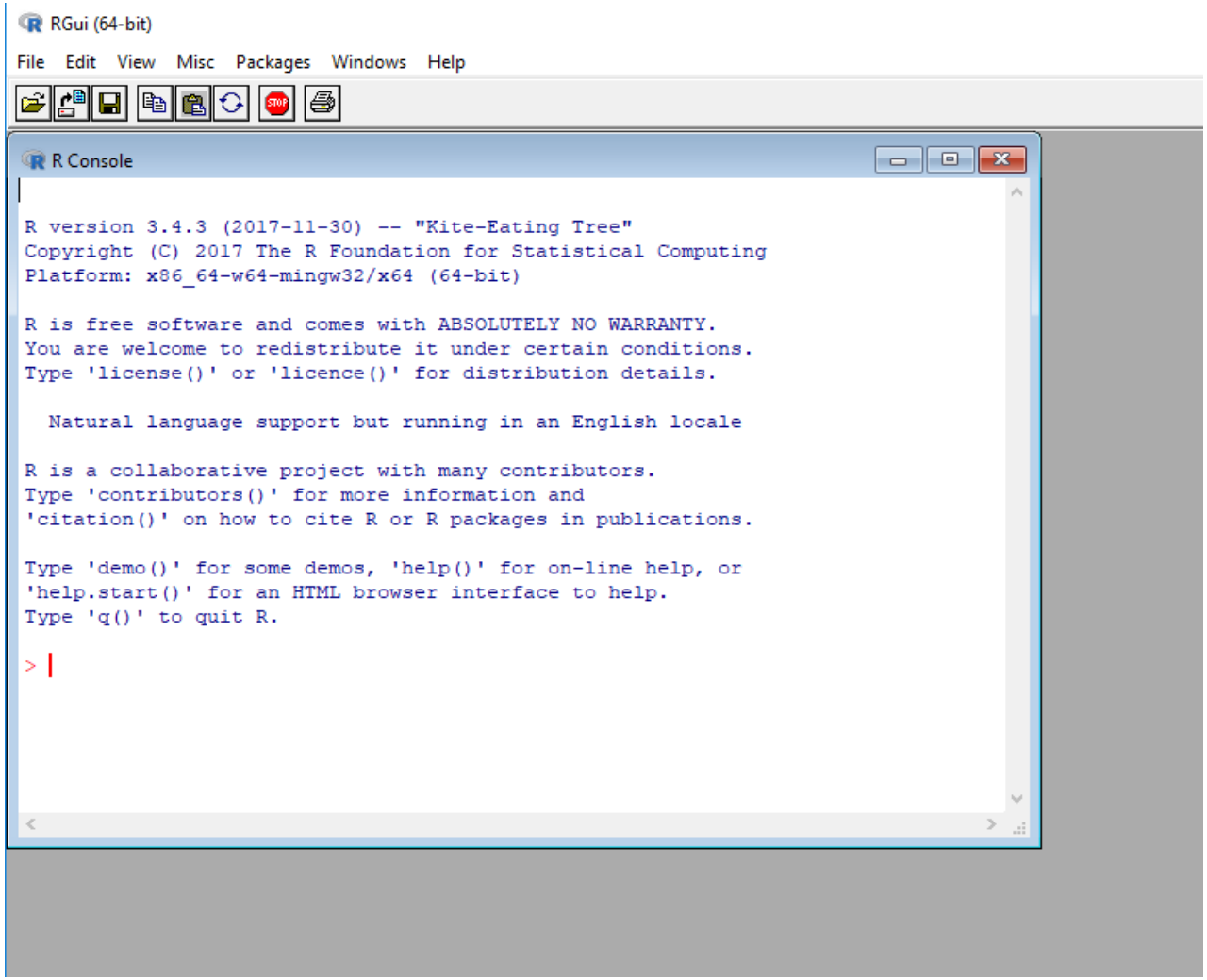
Installation of RStudio

- RStudio can be downloaded at:
  - https://download1.rstudio.org/RStudio-1.1.463.dmg (macOS)
  - https://download1.rstudio.org/RStudio-1.1.463.exe (Windows)

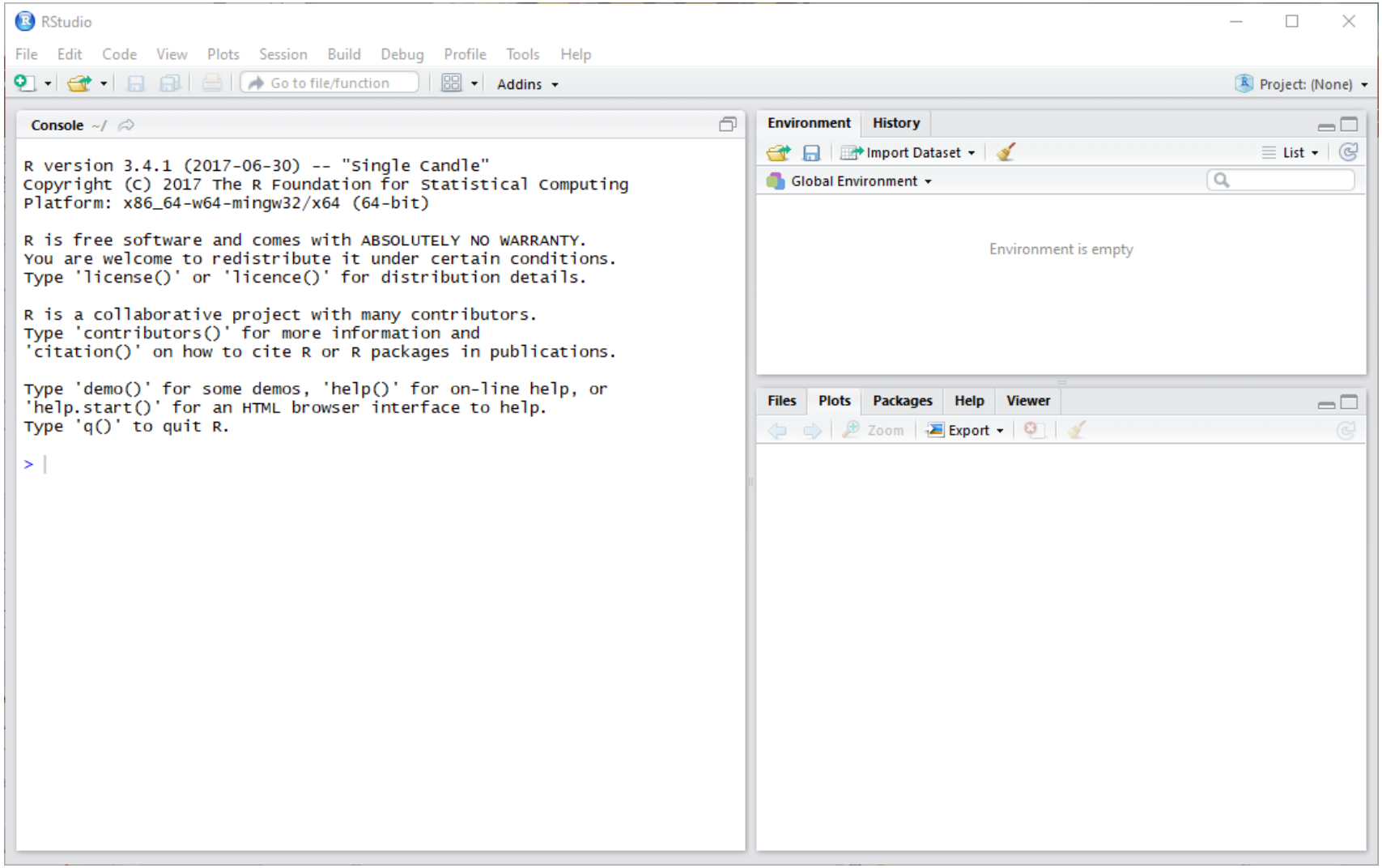
Accessing R

- 3 Ways to access R in Windows / macOS:
  i. via **R Console** (come together with R environment)
  ii. via **Command Prompt** (Win) / **Terminal** (macOS)
    - Type "R" in the command prompt or terminal and press enter to initialize the R environment.
  iii. via **RStudio**
    - A more organized and integrated UI

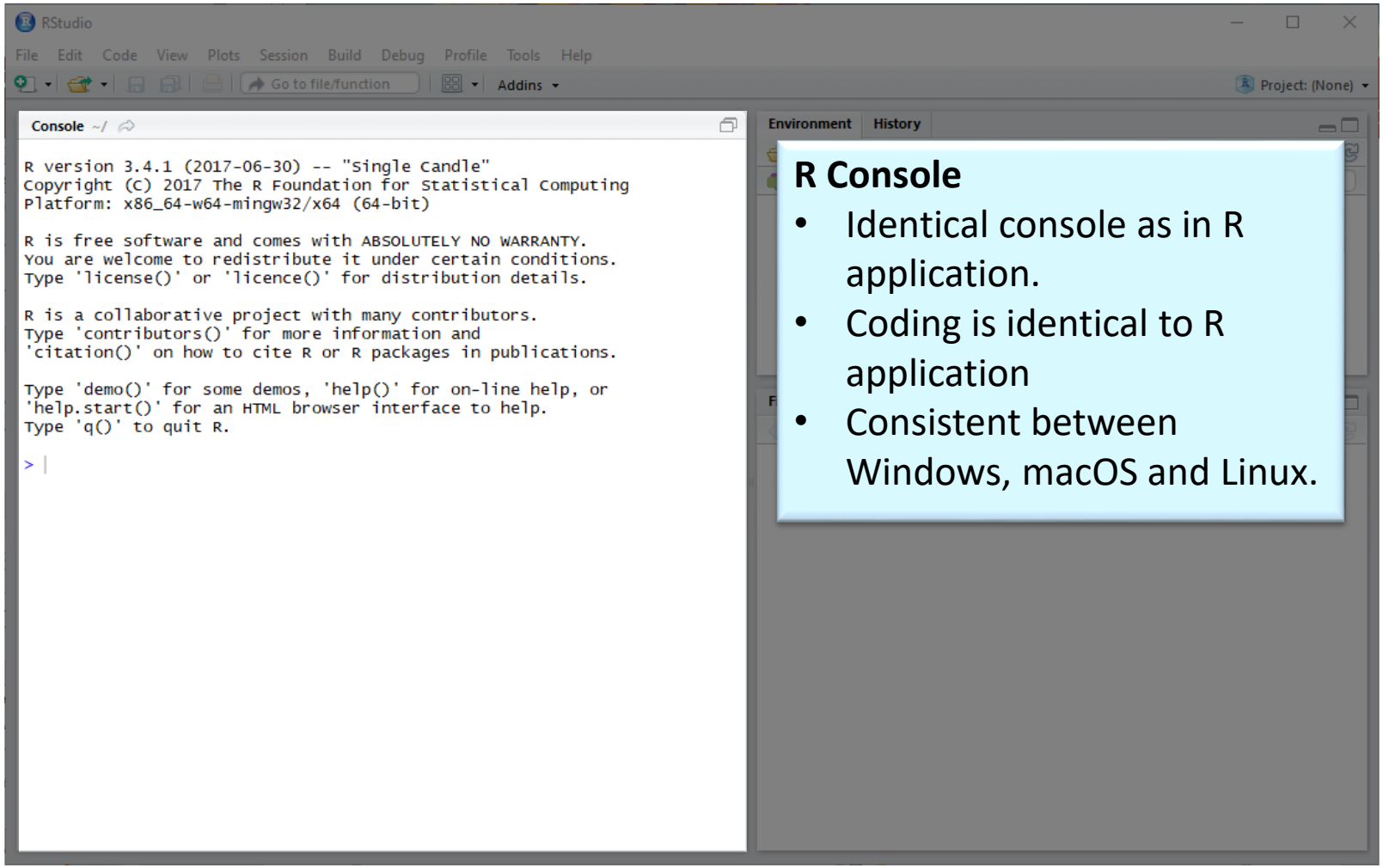
R Console Interface



RStudio Interface



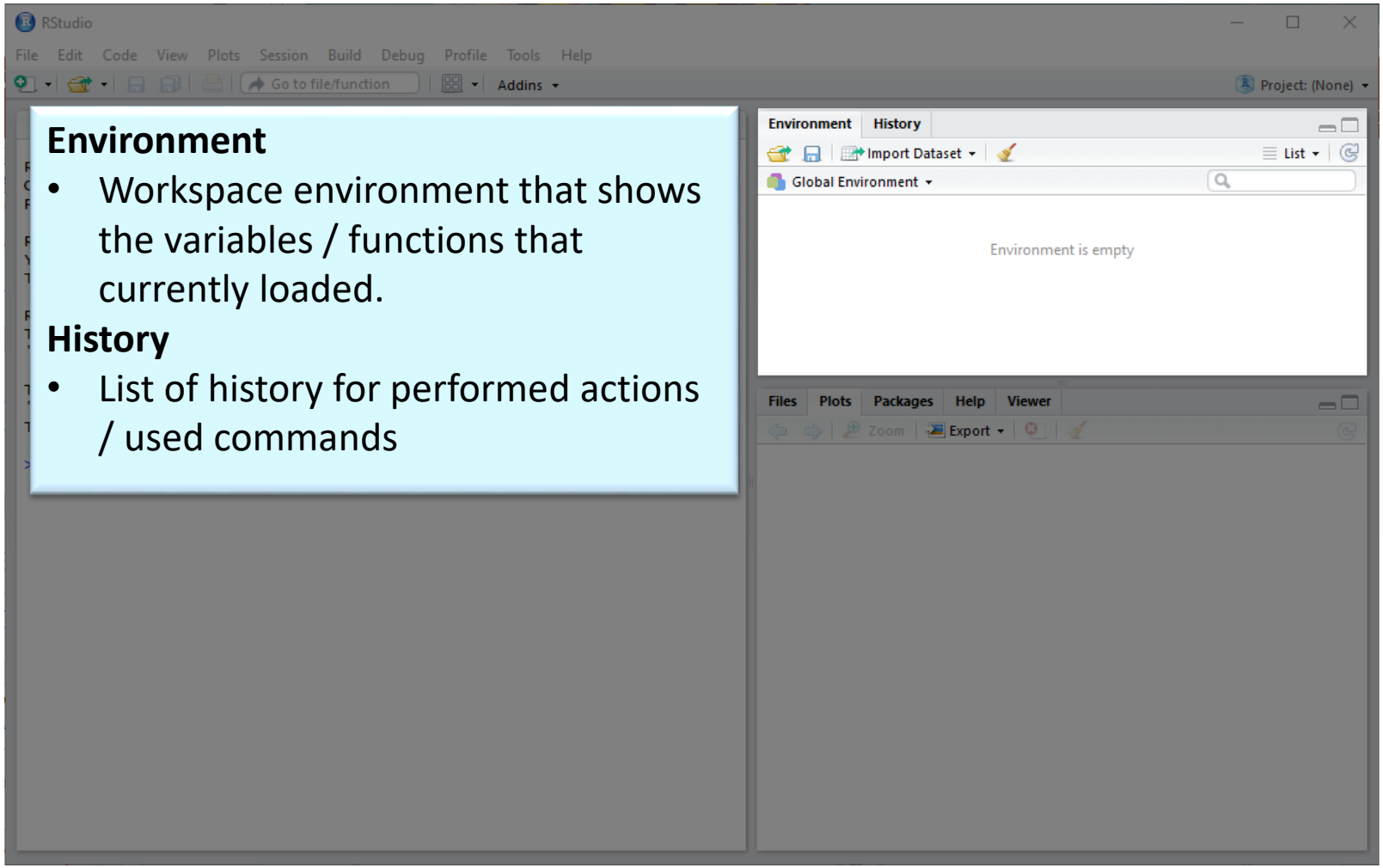
RStudio Interface



R Console

- Identical console as in R application.
- Coding is identical to R application
- Consistent between Windows, macOS and Linux.

RStudio Interface



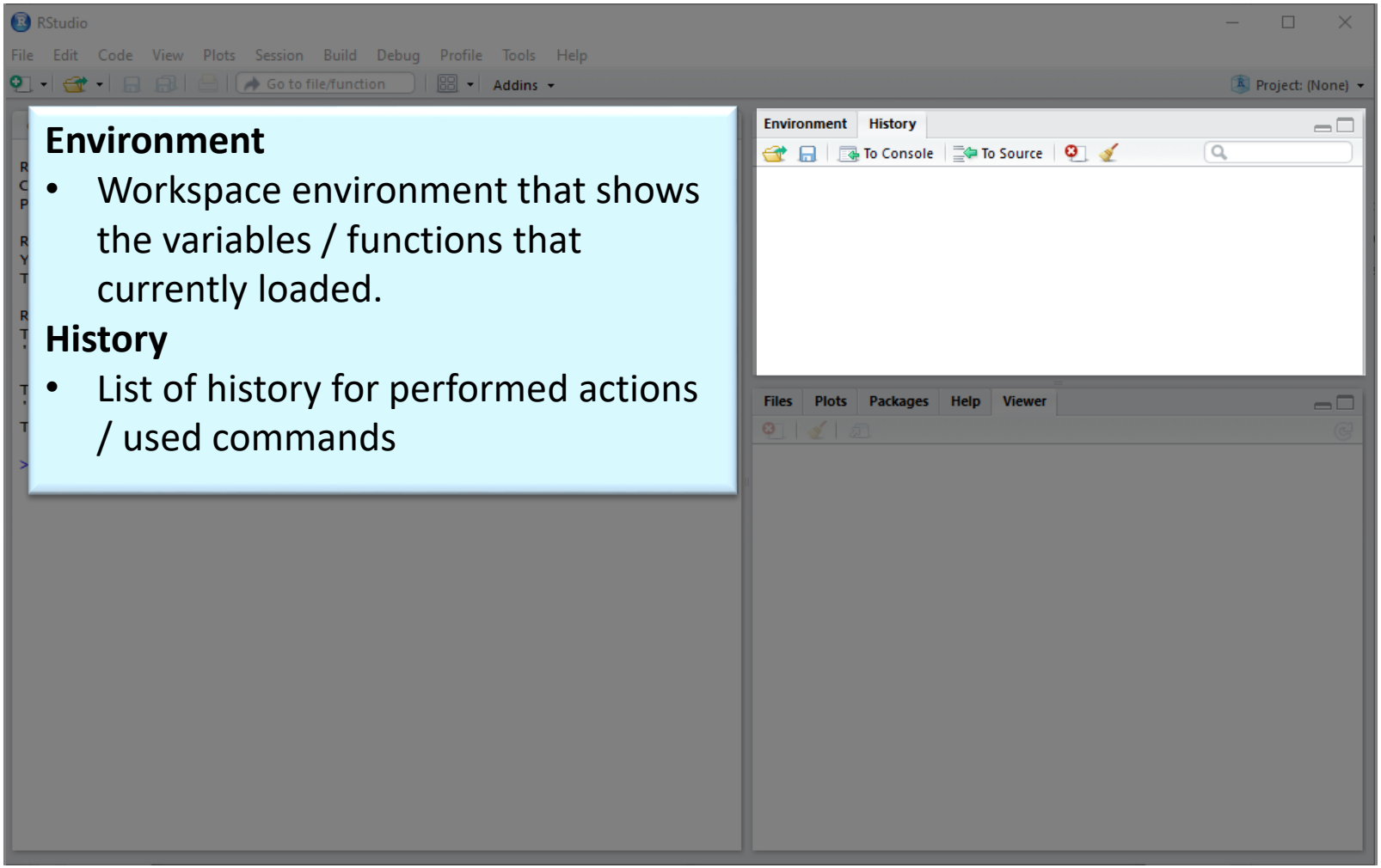
Environment

- Workspace environment that shows the variables / functions that currently loaded.

History

- List of history for performed actions / used commands

RStudio Interface



Environment

- Workspace environment that shows the variables / functions that currently loaded.

History

- List of history for performed actions / used commands

RStudio Interface

Files

- View the folder structure and the contents of the current working directory.

Plots

- Display and settings of the charts / graph

Packages

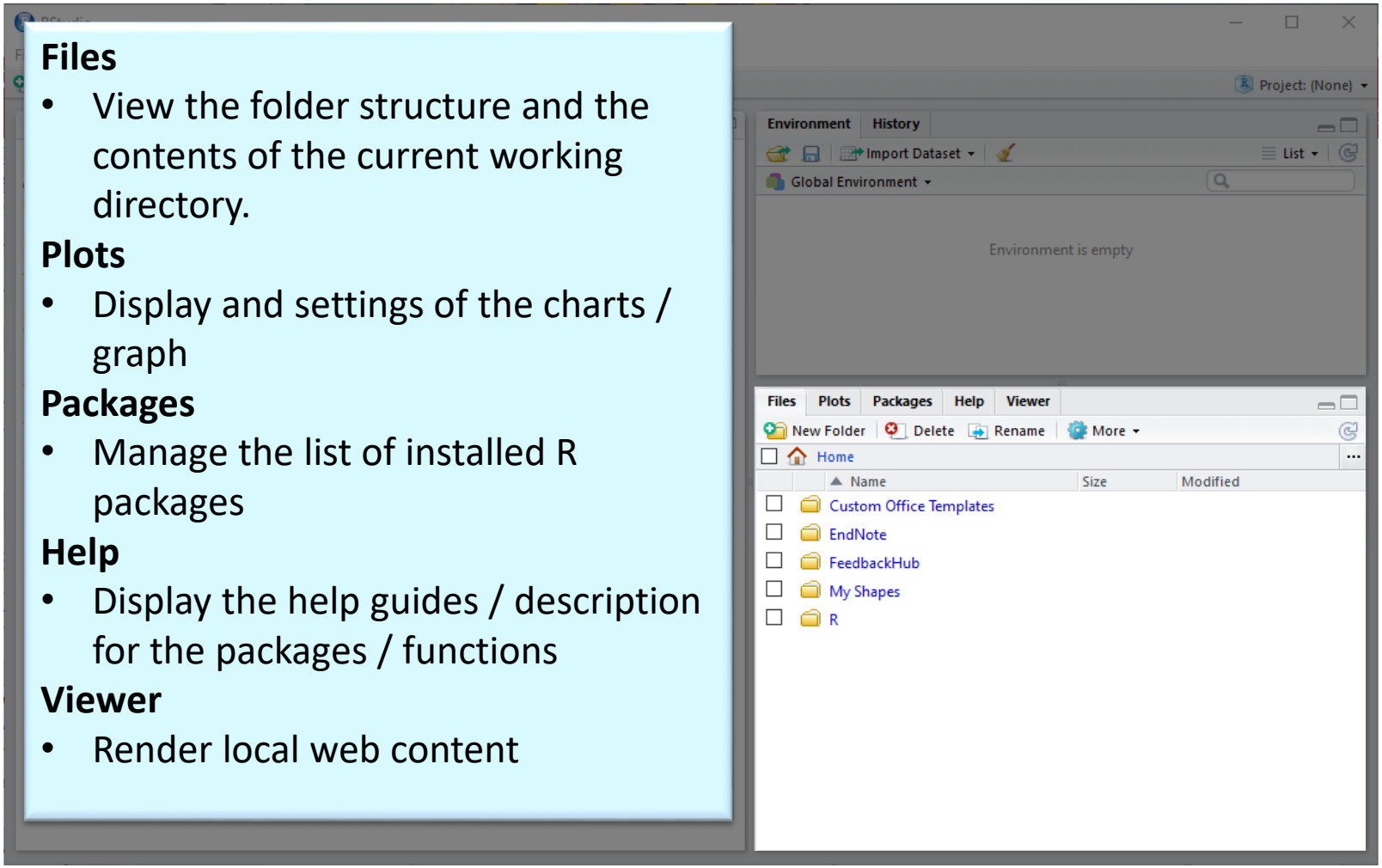
- Manage the list of installed R packages

Help

- Display the help guides / description for the packages / functions

Viewer

- Render local web content



RStudio Interface

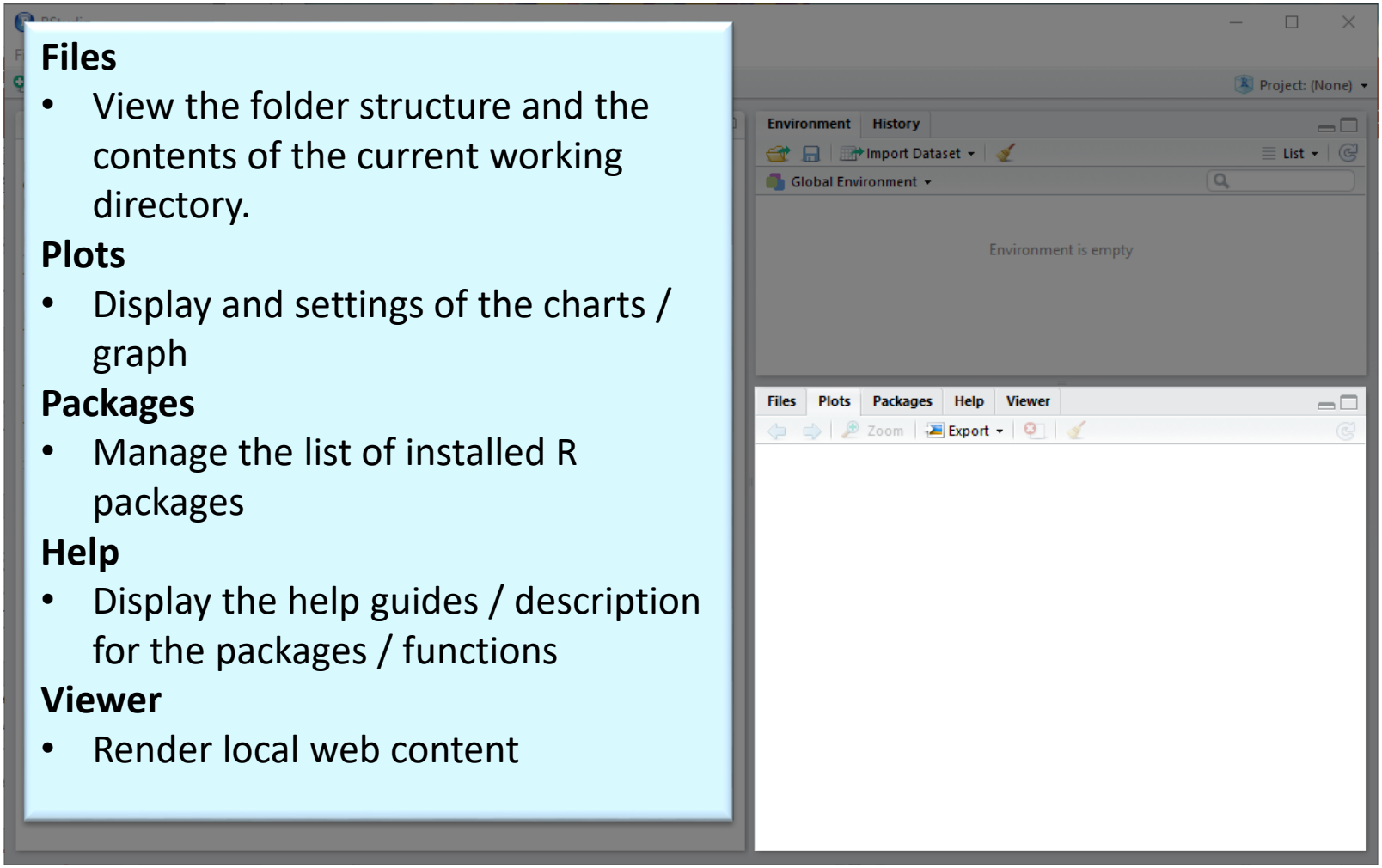
Files
- View the folder structure and the contents of the current working directory.

Plots
- Display and settings of the charts / graph

Packages
- Manage the list of installed R packages

Help
- Display the help guides / description for the packages / functions

Viewer
- Render local web content



RStudio Interface

Files

- View the folder structure and the contents of the current working directory.

Plots

- Display and settings of the charts / graph

Packages

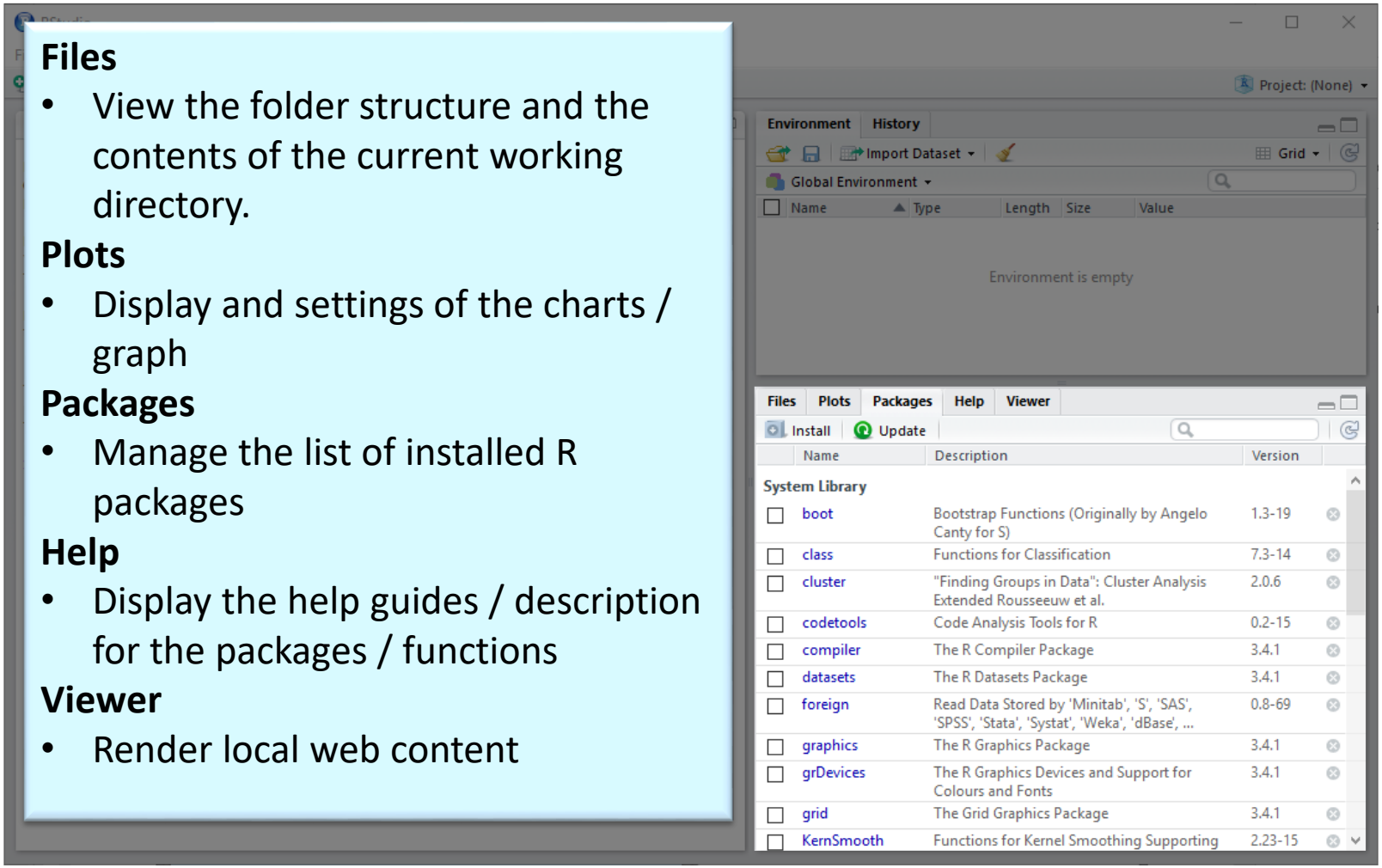
- Manage the list of installed R packages

Help

- Display the help guides / description for the packages / functions

Viewer

- Render local web content



RStudio Interface

Files

- View the folder structure and the contents of the current working directory.

Plots

- Display and settings of the charts / graph

Packages

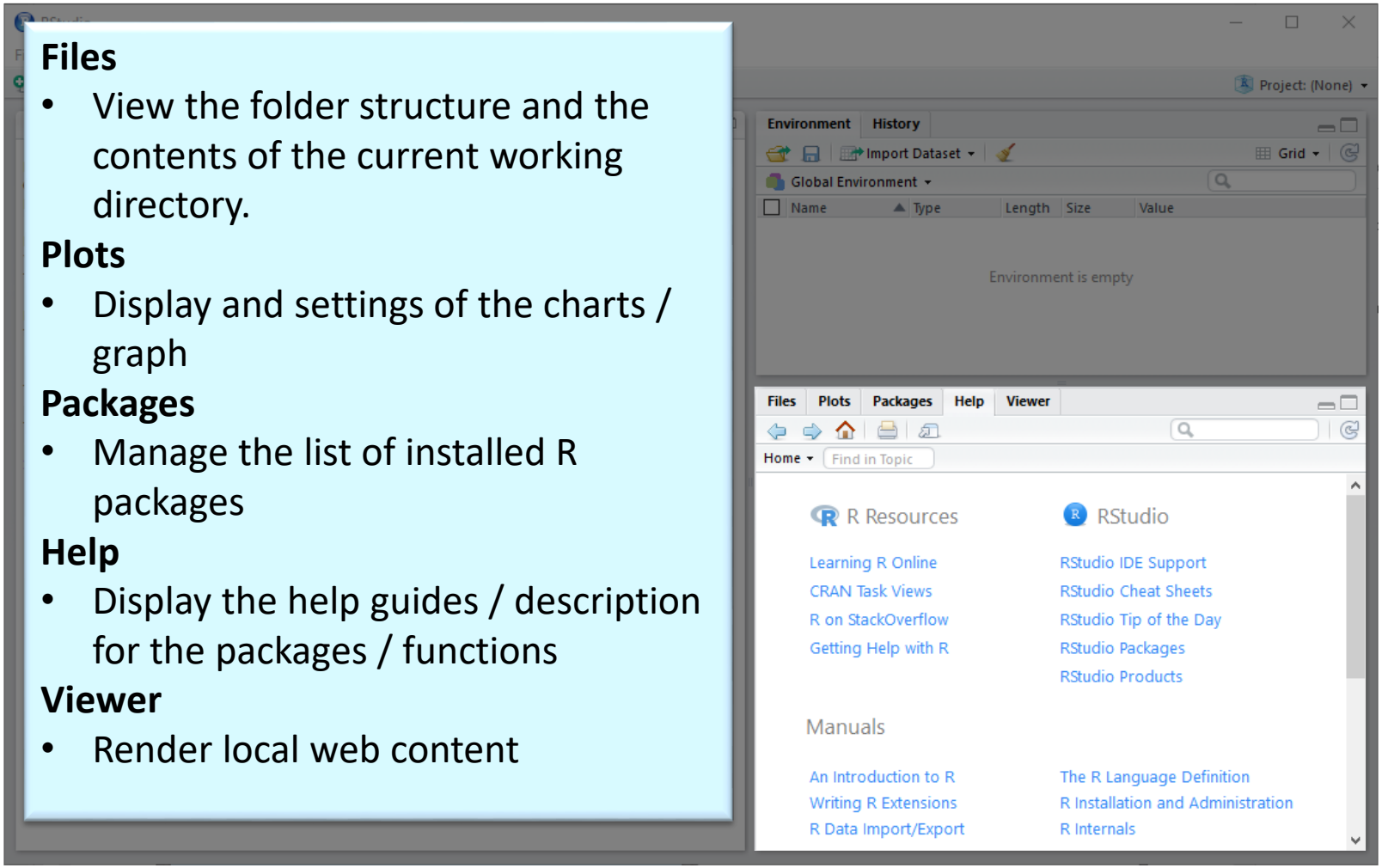
- Manage the list of installed R packages

Help

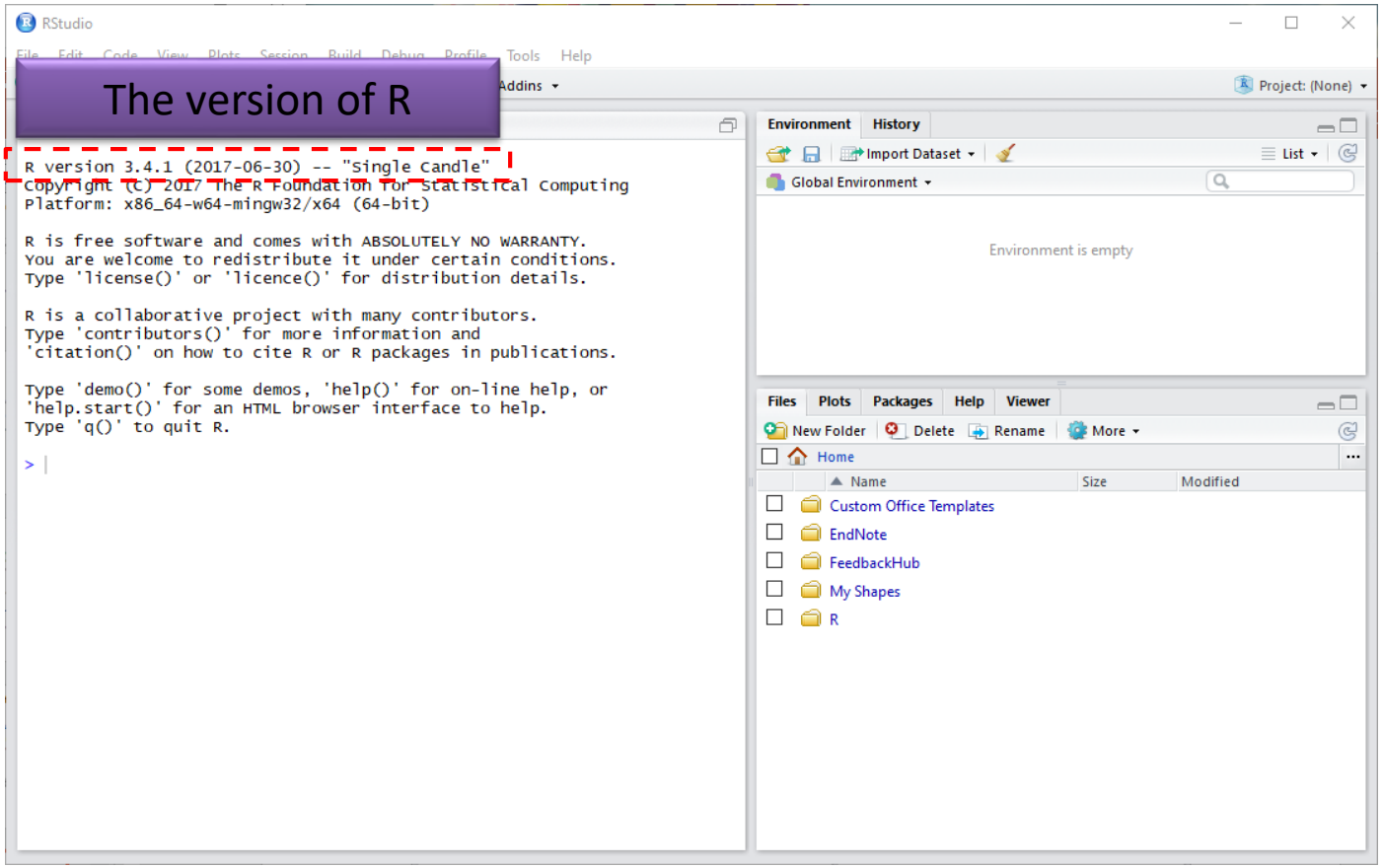
- Display the help guides / description for the packages / functions

Viewer

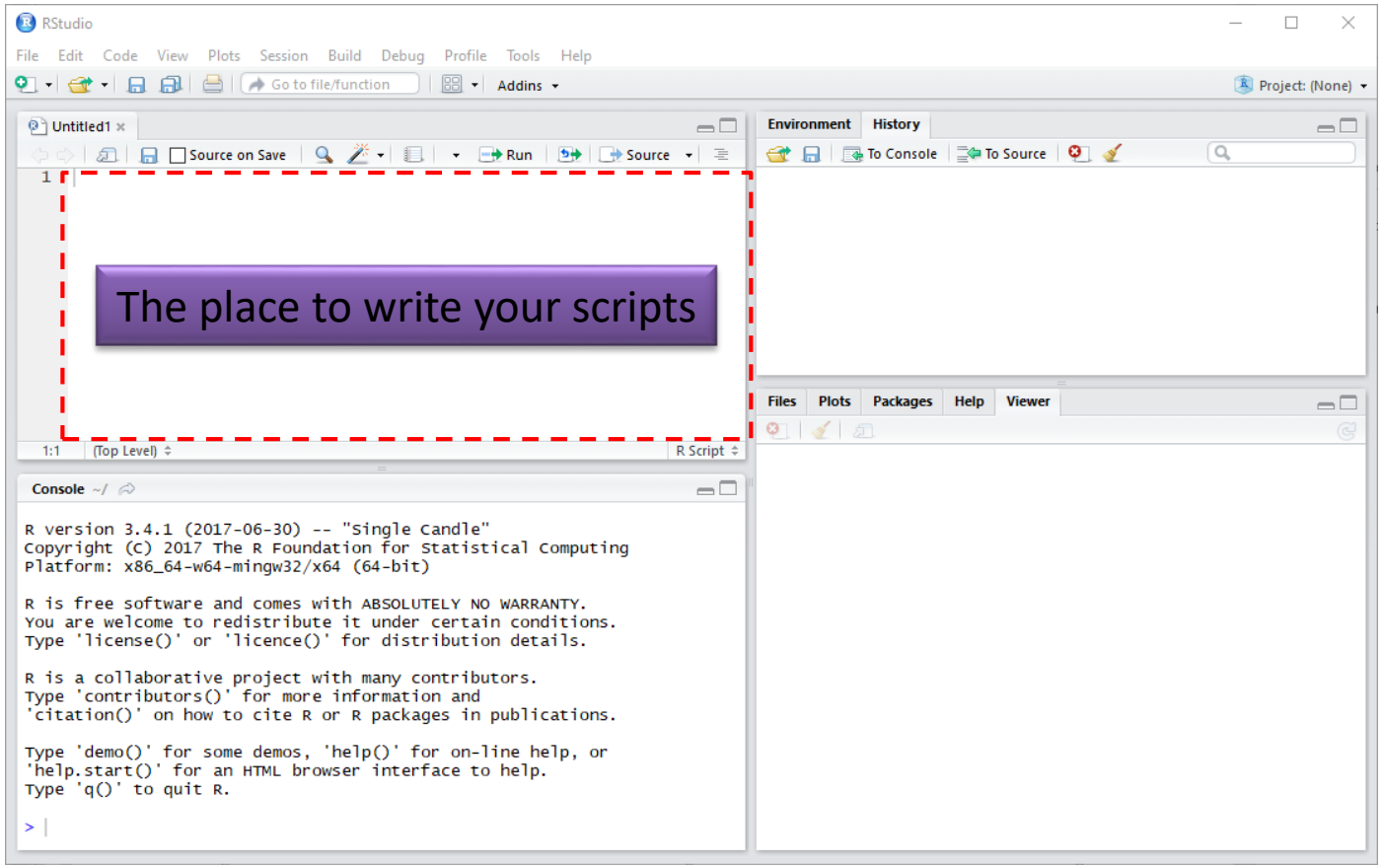
- Render local web content



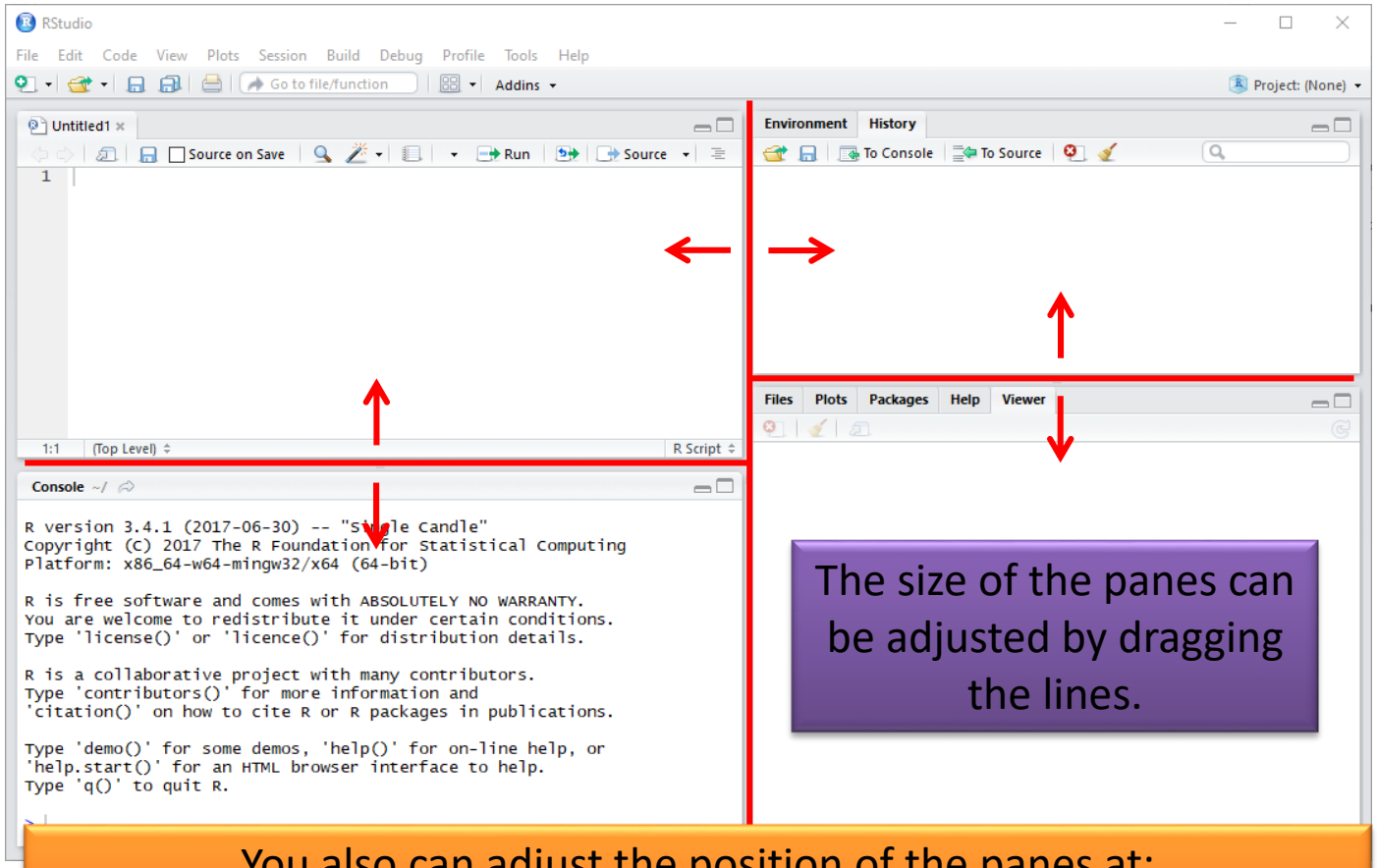
Getting Started with the R Environment



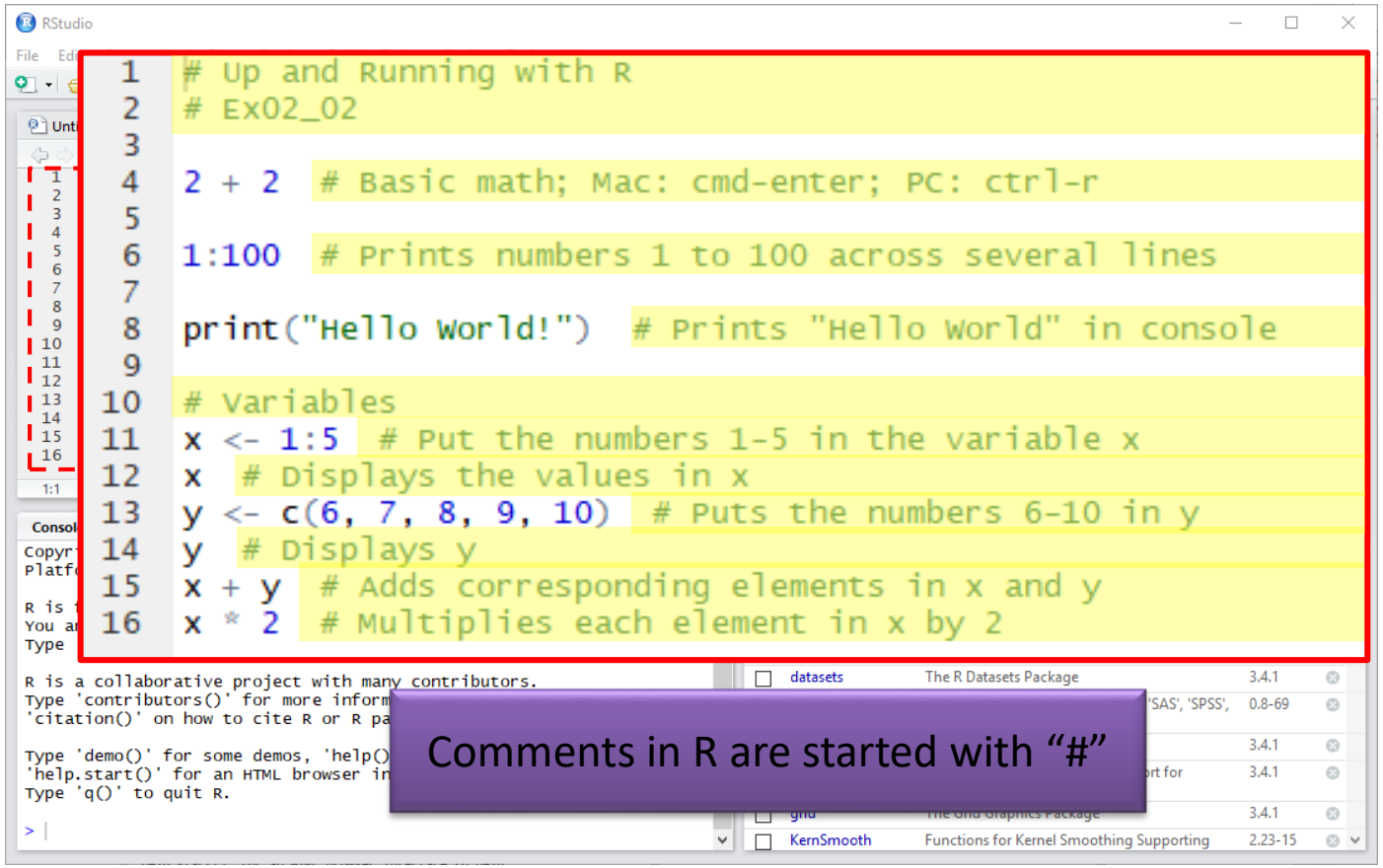
Getting Started with the R Environment



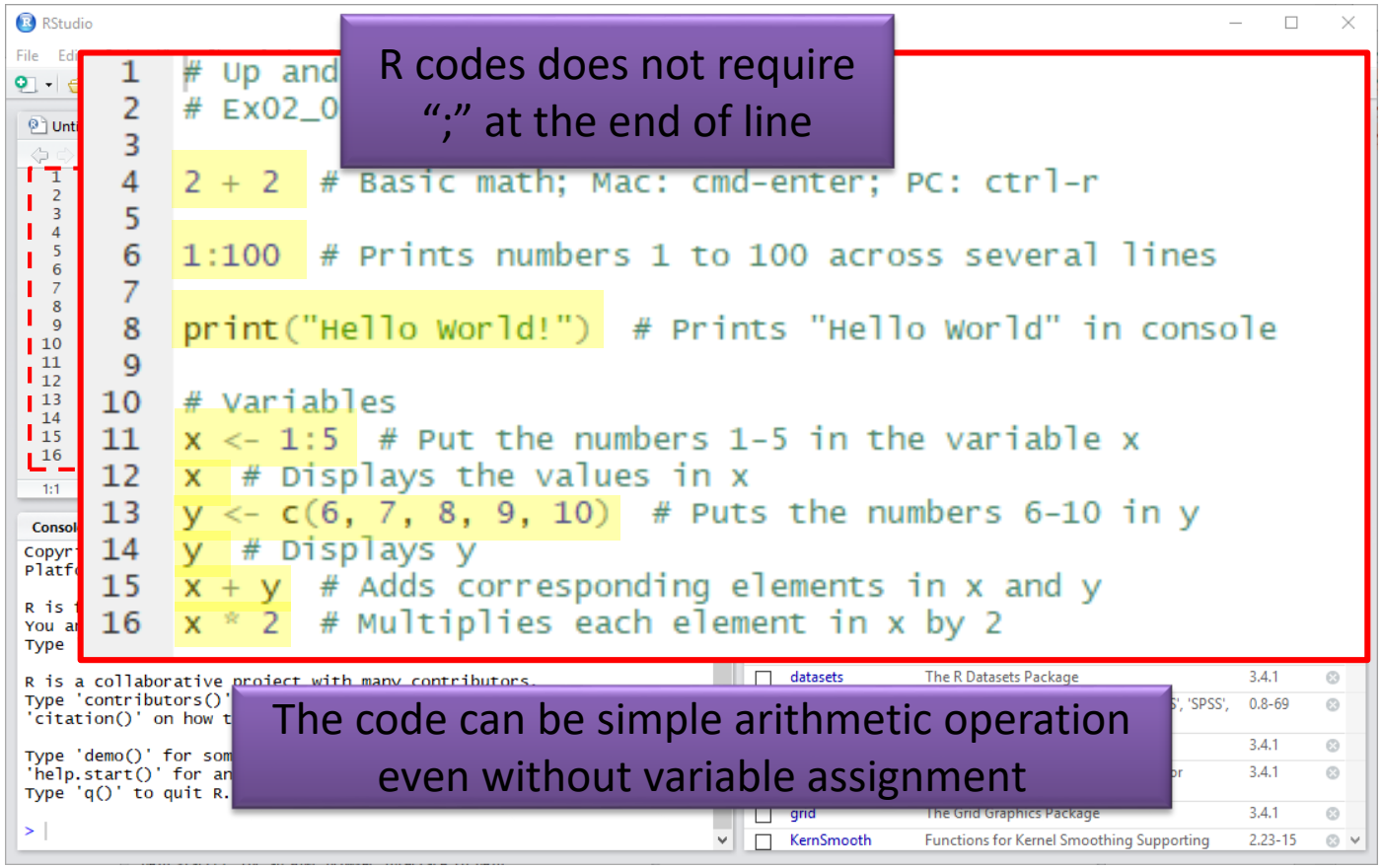
Getting Started with the R Environment



Getting Started with the R Environment

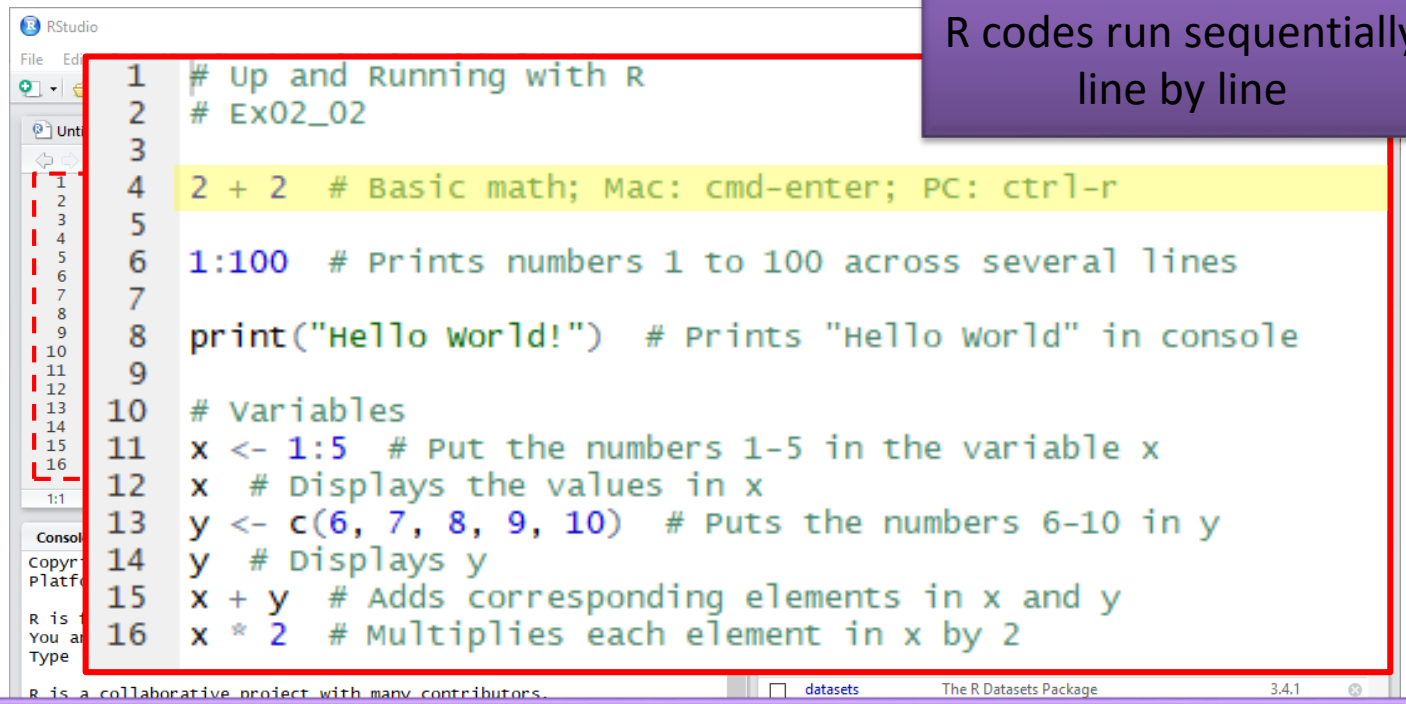


Getting Started with the R Environment



Getting Started with the R Environment

R codes run sequentially line by line

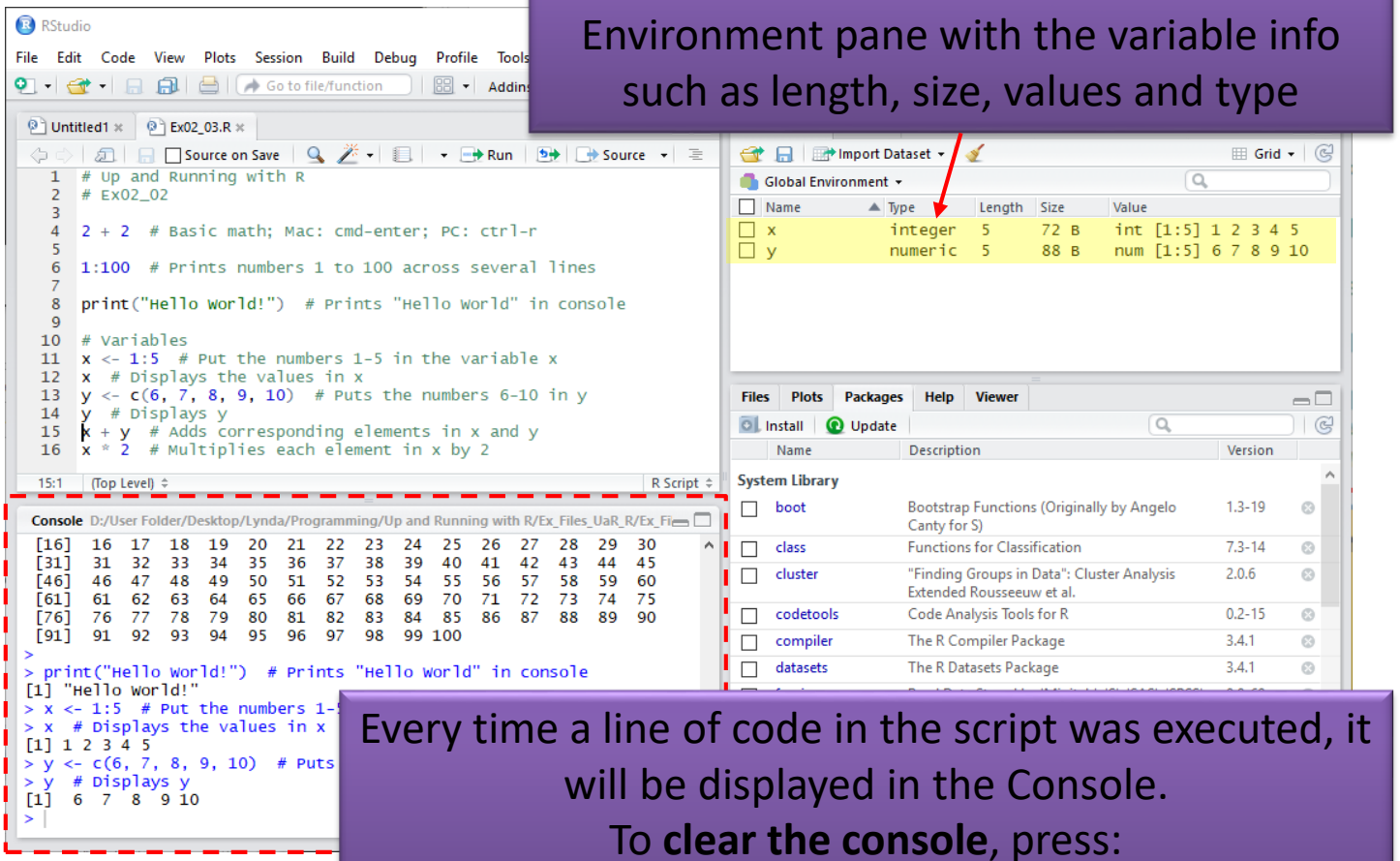


You can run a single line of code by putting a cursor at the start of the line, or run specific part of code by highlighting the code,
Then, press:

Command + **Enter** (macOS) / **CTRL** + **Enter** (Windows)

Getting Started with the R Environment

Variables created will be listed in the Environment pane with the variable info such as length, size, values and type



Every time a line of code in the script was executed, it will be displayed in the Console.

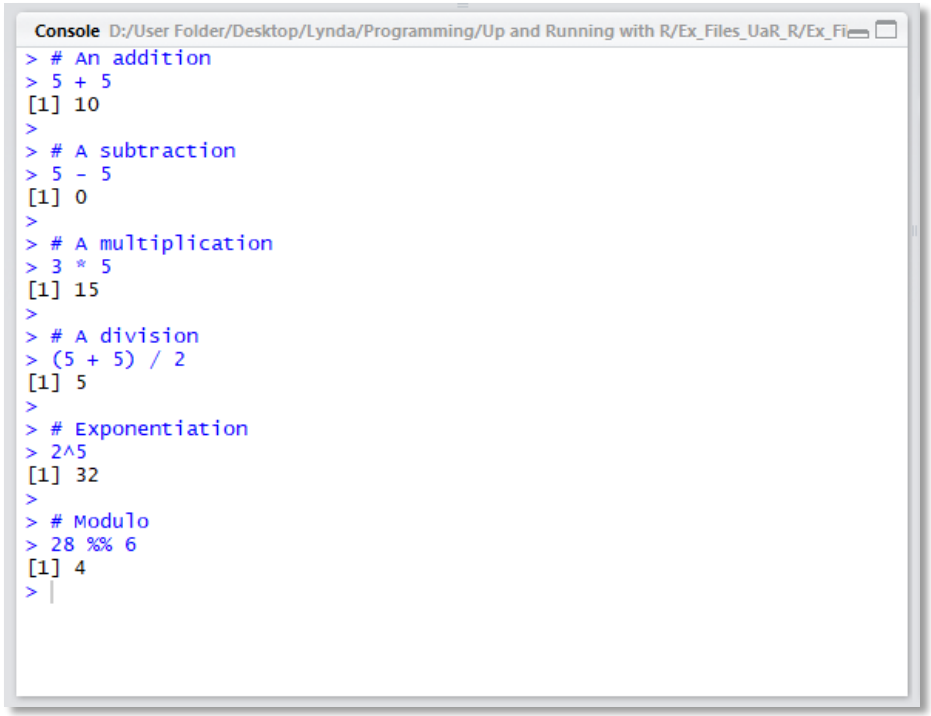
To **clear the console**, press:

Command + **L** (macOS) / **CTRL** + **L** (Windows)

R BASICS

Arithmetic with R

- In the most basic form, R can be used as a simple calculator.
- Using following arithmetic operators:
  - Addition: +
  - Subtraction: -
  - Multiplication: *
  - Division: /
  - Exponentiation: ^
  - Modulo: %%



Variable Assignment in R

- In R, variable assignment is done using either “<-” or “=“
- For example: **“x<-4”** is equivalent to **“x=4”**
- However, some style guide such as Google R Style Guide prefers the use of “<-” instead of “=“

Google's R Style Guide

R is a high-level programming language used primarily for statistical computing and graphics. The goal of the R Programming Style Guide is to make our R code easier to read, share, and verify. The rules below were designed in collaboration with the entire R user community at Google.

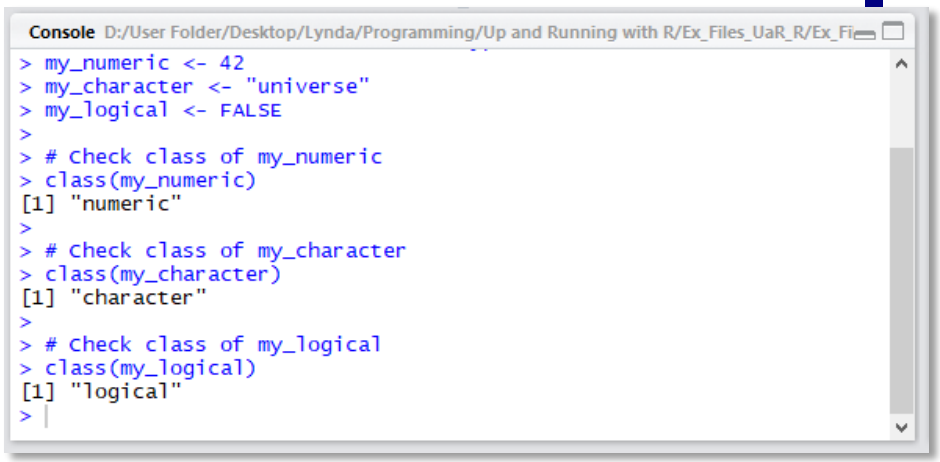
Summary: R Style Rules

1. File Names: end in `.R`
2. Identifiers: `variable.name` (or `variableName`), `FunctionName`, `kConstantName`
3. Line Length: maximum 80 characters
4. Indentation: two spaces, no tabs
5. Spacing
6. Curly Braces: first on same line, last on own line
7. else: Surround else with braces
8. Assignment: use `<-`, not `=`
9. Semicolons: don't use them
10. General Layout and Ordering

https://google.github.io/styleguide/Rguide.xml

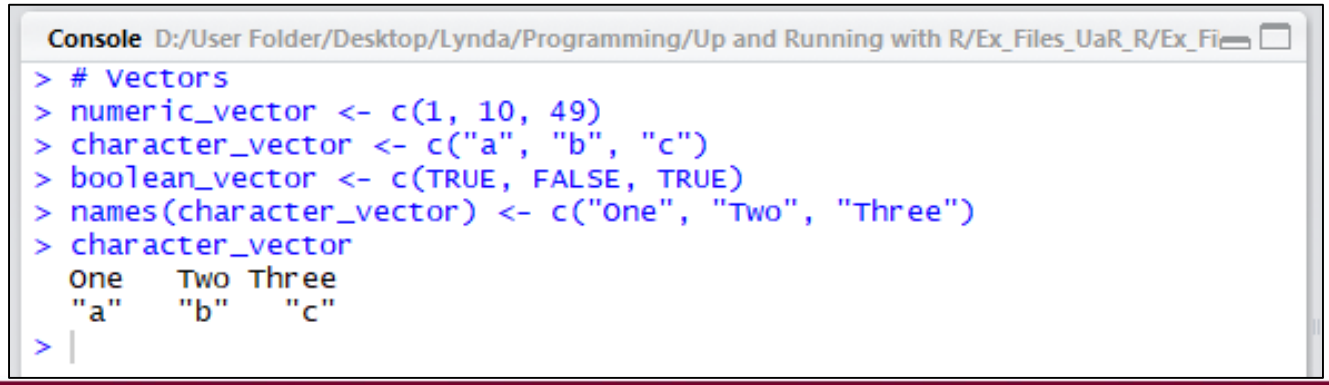
Basic Data Types in R

- Some of the basic data types in R are:
  - **Numeric**: ……1, 10, 15, 1C, 2E,……
  - **Integer**: ……2, 15, 33, 13, 18,……
  - **Logical**: TRUE or FALSE
  - **Characters**: Text or strings
- The data types of a variable can be checked with **class()** function.



Vectors in R

- Vectors are **one-dimension arrays** that can hold numeric, integer, character or logical data.
- In R, vectors are created using function **c()** ,for example:
  - **numeric_vector <- c(1, 10, 49)**
  - **character_vector <- c("a", "b", "c")**
  - **boolean_vector <- c(TRUE, FALSE, TRUE)**
- In R, we can set custom name for the vector. Example:
  - names(character_vector) <- c("One", "Two", "Three")



Arithmetic of Vectors in R

- Vectors can perform **element-wise arithmetic operation**.
- Sum and average of the elements in the vector can be calculated using **sum(*vector*)** and **mean(*vector*)** respectively.
- Vectors can used to do comparison, for example:
  - *vector = c(1,2,3,4,5)*
    vector > 3
    Output:
    FALSE FALSE FALSE TRUE TRUE
- Element in specific position in the vector can be selected using **vector[*position*]**.
  - Bool_value = c(TRUE, FALSE, TRUE, FALSE)
    Vector = c(2, 3, 5, 6)
    Vector[Bool_value]
    Output:
    2 5

Factors in R

- Factors are **categorical variables** that are useful in summary statistics, plots and regressions.
- To create factors in R, use **factor()** function.
- The factor levels can be changed using the **levels()** function.

```
my_vector <- c("L", "S", "L", "M", "M")

# Option 1
my_factor <- factor(my_vector)
levels(my_factor) <- c("Large", "Medium", "Small")

# Option 2
my_factor <- factor(my_vector,
                    levels = c("S", "M", "L"),
                    labels = c("Small", "Medium", "Large"))
```

Matrices in R

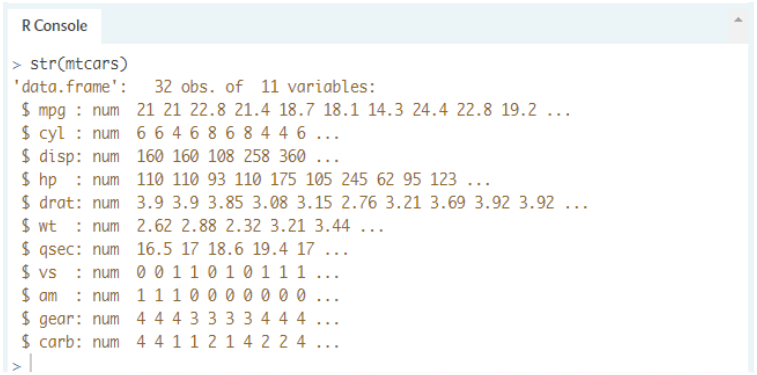
- In R, a matrix is a **collection of elements** of the same data type (numeric, character, or logical) arranged into a fixed number of **rows** and **columns**.
- Since we are only working with rows and columns, a matrix is called **two-dimensional**.
- You can construct a matrix in R with the **matrix()** function. Consider the following example:
  - matrix(1:9, byrow = TRUE, nrow = 3)
- The row and column of the matrix can be named using **rownames()** and **colnames()** respectively.
  - Example: colnames(*matrix*) <- names

Operations in Matrices

- We can calculate the column sum and row sum using **colSums()** and **rowSums()** respectively.
  - Example: colSums(matrix)
- Additional data can be bind into the matrix using **cbind()** and **rbind()**.
  - cbind(matrix_A, matrix_B): add matrix_B into the "right" side of matrix_A
  - rbind(matrix_A, matrix_B): add matrix_B into the "bottom" side of matrix_A
- Element in specific position in the matrix can be selected using **matrix[row_position, col_position]**.

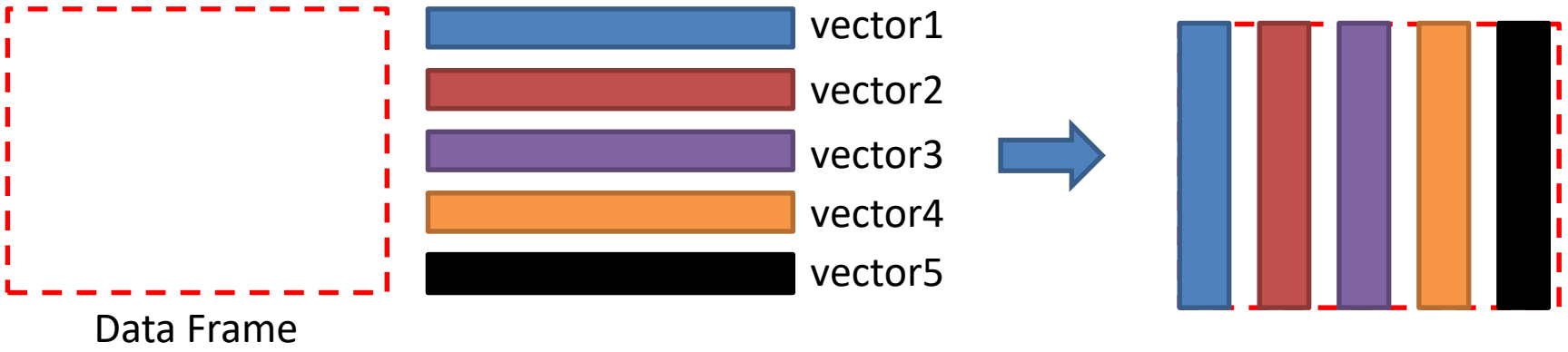
Data Frame in R

- In R, data in matrix should consists of same data type. However, if the data consists of **different data types**, a data frame is more suitable for this case.
- When dealing with huge data frame, we can use **head(data frame)** or **tail(data frame)** to show only the first few observations or last few observations of the data frame respectively.
- In order to inspect the structure of the data frame, we can use the function **str()**



Creating a Data Frame

- Data frame can be constructed with the **data.frame()** function.
  - Example: data.frame(*vector1, vector2, vector3, vector4, vector 5*)
- Different vectors will be passed as argument and they will become the **different columns of your data frame**.
- Because every column has the same length, the vectors you pass must also have the **same length**.



Accessing Data Frame

- Data frame can be access via **dataframe[row position, column position], dataframe[row position, column name]**
- Column of the data frame also can be assess by using "$"
  - For example: **dataframe$column1**
- Subset of the data frame can be obtained through function **subset(dataframe, subset=condition)**
- The data frame can be sorted using **order()** function.
  - For example: **order(dataframe$column)**

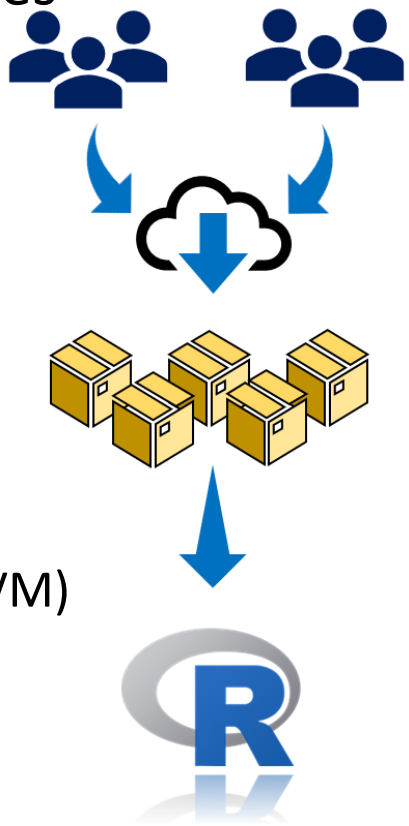
Characters Data in R

- Characters data in R often use **double quotes**.
- **Concatenation** of strings can be done using **paste()** function. For example: paste(string1, string2, sep="")
- To print/display output on the console window, use **print()** function.

PACKAGES IN R

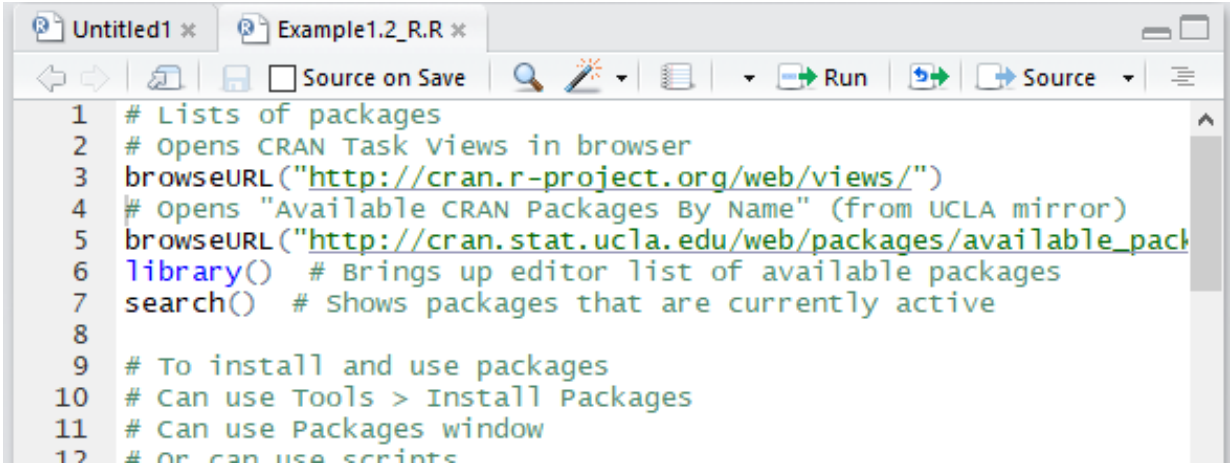
R Packages

- R is highly extensible with widely available packages contributed by community.
- Easy sharing of scripts.
- Comprehensive R Archive Network (CRAN)
- Currently lists over 10000 free packages for R
- Example of packages such as:
  - ggplot2 – for advanced visualization of data
  - e1071 – standard library for support vector machine (SVM)
- To manage R packages, there are two ways:
  - via R Console
  - via RStudio



List available R Packages

- Line 3 code will open up the URL in web browser which shows a large list of categories of packages that are available.
- Line 5 code will shows a long list of packages by name.
- To show the installed packages in the machine, use **library()**
- To show the current active packages, use **search()**



Install R Packages

- There are several ways to **install** R packages in RStudio.
  - Method 1: Using function install.packages()
    - For example: install.packages("package_name")
  - Method 2: Install package through RStudio GUI (online/manual)
    - Can be accessed through:
      - Package pane at the bottom right and click install.
      - Tools menu > Install Packages
- To **load** the installed packages, use:
  - library("*package_name*")
  - require("*package_name*")
- To bring up the **documentation** of the package:
  - library(help="*package_name*")

Update, Detach and Delete R Packages

- To update packages:
  - Use function: **update.packages()**
  - Or through Tools > Checks for Updates
- To unload/detach an active package:
  - Use function: **detach("package:*package_name*", unload=TRUE)**
  - Untick the function that want to detach at Package pane.
- To delete an installed package:
  - Use function: **remove.packages("*package_name*", lib="~/R/win-library/3.4")**
  - Click the "x" at the right of the list of installed package in the Package pane.

SEARCH HELP IN R

Vignettes in R

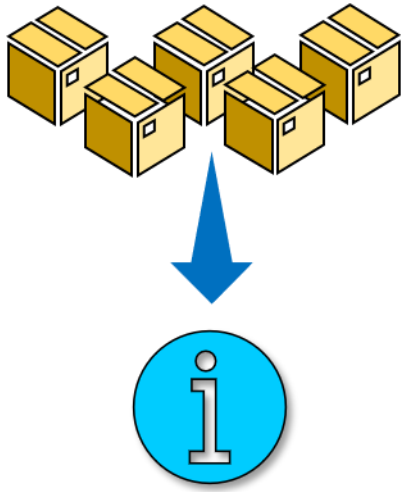
- Most of the R packages have a list of vignettes, which consist of the examples and detail description regarding the package.
- To bring up the list of vignettes of the package:

  vignette(package = "*package_name*")
- To show an interactive list of vignettes of the package in HTML through browser:

  browseVignettes(package = "*package_name*")
- To show all vignettes for all available packages:

  vignette()

  browseVignettes()



FUNCTIONS IN R

Define a function

- To create a “trigger” for a series of action in a more manageable way.
- Note: the R script must be loaded first before the function can be called in the R Console.

Script.R

```
myFunctName <- function() {
        print("Hello")
}

Function name                    Parameters
myFunct2Name <- function(param1, param2)
{
        print(param1+param2)
}

Function name
myFunct3Name <- function(param1) {
        myFunction2Name(param1, 10)
        print(param1)
}
```

R Console

```
> myFunctName()
[1] Hello
> myFunct2Name(15, 4)
[1] 19
> myFunct3Name(25)
[1] 35
[1] 25
```

CONTROL STRUCTURE IN R

Control Structures in R

- If……Else in R

```
if(condition) {
        #statement
}
```

- For loop in R

```
for(x in 1: 10) {
        print(x)
}
```

- While loop in R

```
x <- 1;
while(x<11) {
        print(x)
        x = x+1;
}
```

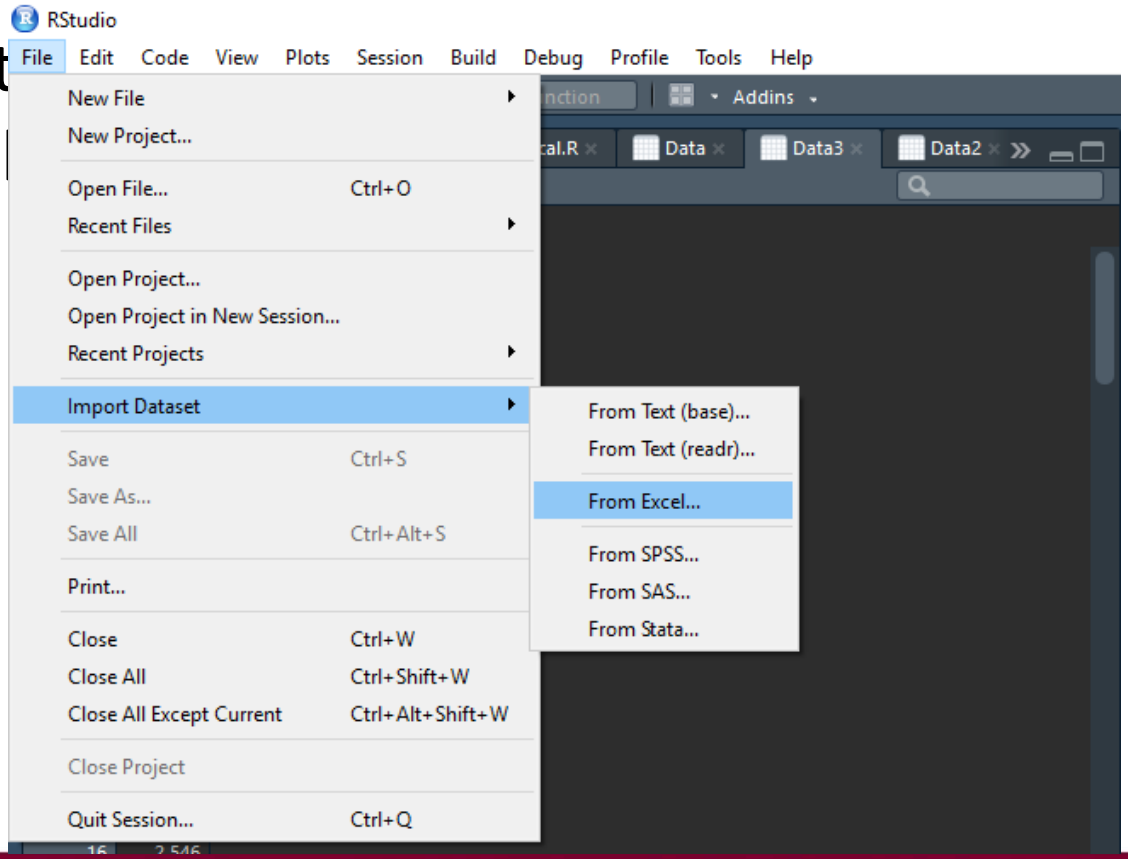
IMPORT/EXPORT IN R

Save/Load and Import/Export

- In R, data can be save into R binary data (.Rdata extension), which only readable by R.
- This can be done by **save()** function.
  - For example: save(*variable*, file="*path/filename.Rdata*")
  - The data saved in Rdata format can be loaded into R using **load()**
- In R, data also can be exported to a Excel-readable format (.csv) using **write.csv()**
  - For example: write.csv(*dataframe, "path/filename.csv"*)
  - The CSV file can be imported into R using **read.csv()**

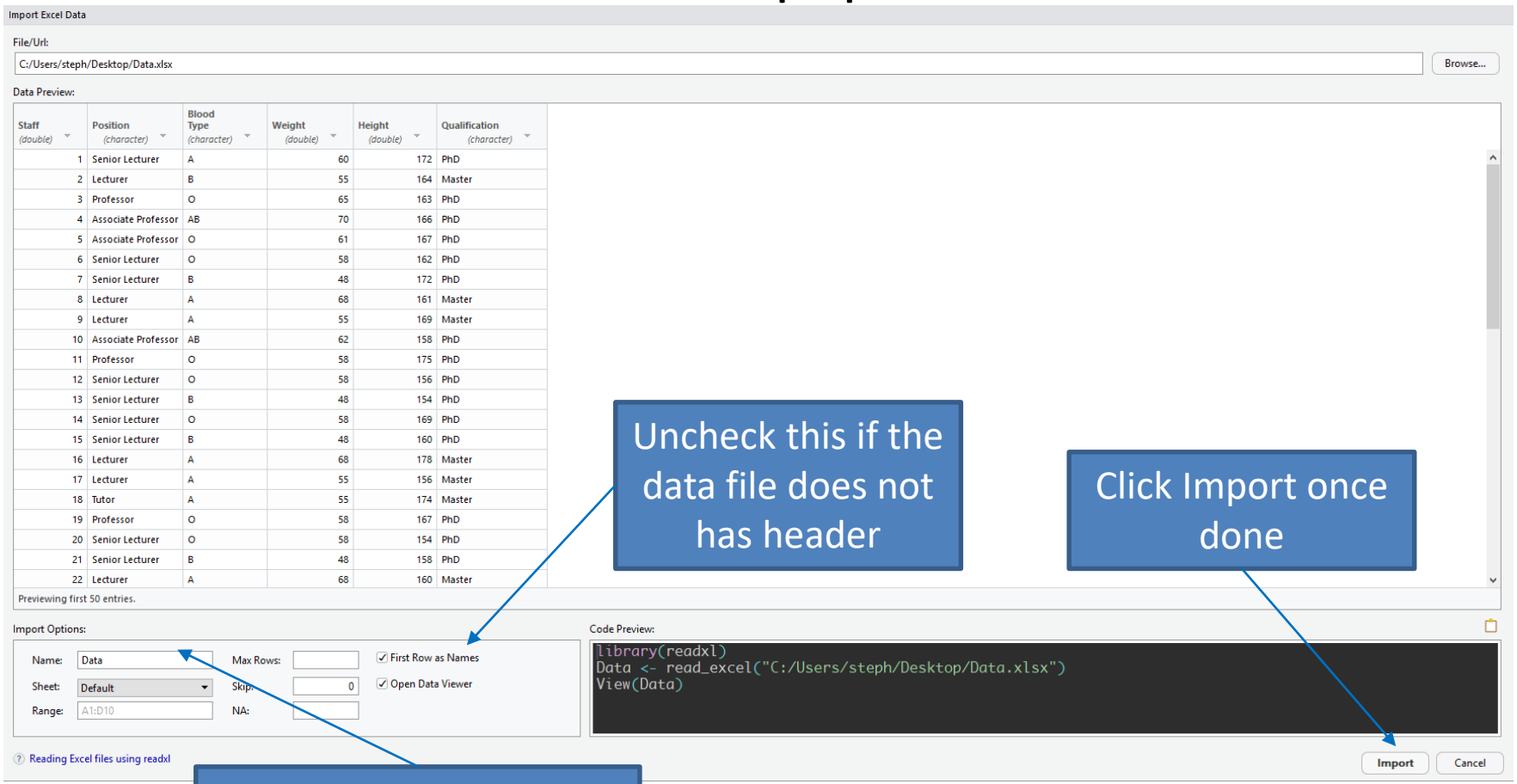
Import from Excel

- In R, Excel file can be imported by navigating to the option pane: File→Import Dataset→From Excel
- If first
  it wi
  download
  install a package.



Import from Excel

- File selection window will pop out as follow:



STATISTICS WITH R

Presenting Categorical Data

Frequency Distribution

- To organize the data using table() function in R
- Example:

	A	B	C	D	E	F
1	Staff	Position	Blood Type	Weight	Height	Qualification
2	1	Senior Lecturer	A	60	172	PhD
3	2	Lecturer	B	55	164	Master
4	3	Professor	O	65	163	PhD
5	4	Associate Professor	AB	70	166	PhD
6	5	Associate Professor	O	61	167	PhD
7	6	Senior Lecturer	O	58	162	PhD
8	7	Senior Lecturer	B	48	172	PhD
9	8	Lecturer	A	68	161	Master
10	9	Lecturer	A	55	169	Master
11	10	Associate Professor	AB	62	158	PhD
12	11	Professor	O	58	175	PhD
13	12	Senior Lecturer	O	58	156	PhD
14	13	Senior Lecturer	B	48	154	PhD
15	14	Senior Lecturer	O	58	169	PhD
16	15	Senior Lecturer	B	48	160	PhD
17	16	Lecturer	A	68	178	Master
18	17	Lecturer	A	55	156	Master
19	18	Tutor	A	55	174	Master
20	19	Professor	O	58	167	PhD
21	20	Senior Lecturer	O	58	154	PhD
22	21	Senior Lecturer	B	48	158	PhD

Academic staff info: 140 staffs

Steps to analyze the data:

1. Import the data into R
2. Extract frequency distribution for desired information
3. Plot the frequency distribution in a bar plot

Presenting Categorical Data

Frequency Distribution

- **STEP 1:** At the menu bar of Rstudio, go to: File → Import Data → From Excel select the data file and import into the R environment.
- **STEP 2:** Extract the frequency distribution information using following command:

```
staffPosition.freq <- table(Data[,2])
```

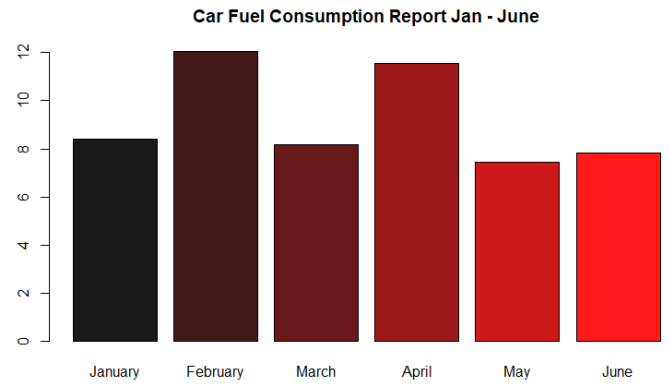
- **STEP 3:** Plot the frequency distribution using bar plot

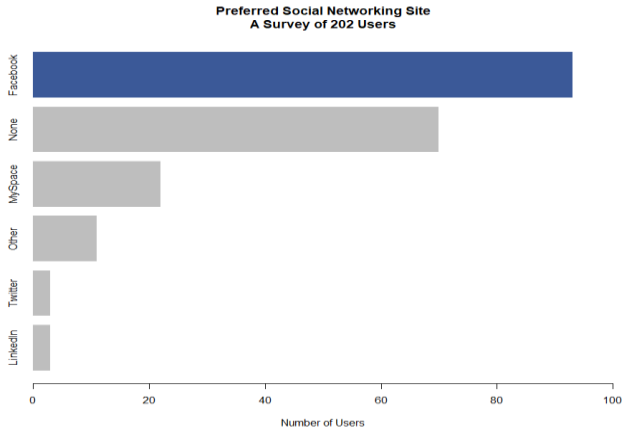
```
barplot(staffPosition.freq)
```

	A	B	C	
1	Staff	Position	Blood Type	We
2	1	Senior Lecturer	A	6
3	2	Lecturer	B	5
4	3	Professor	O	6
5	4	Associate Professor	AB	7
6	5	Associate Professor	O	6
7	6	Senior Lecturer	O	5
8	7	Senior Lecturer	B	4
9	8	Lecturer	A	6
10	9	Lecturer	A	5
11	10	Associate Professor	AB	6
12	11	Professor	O	5
13	12	Senior Lecturer	O	5
14	13	Senior Lecturer	B	4
15	14	Senior Lecturer	O	5
16	15	Senior Lecturer	B	4
17	16	Lecturer	A	6
18	17	Lecturer	A	5
19	18	Tutor	A	5
20	19	Professor	O	5
21	20	Senior Lecturer	O	5
22	21	Senior Lecturer	B	4

Bar Plot in R

- In R, bar chart / barplot is plotted using **barplot()** function.
- **barplot()** is a built-in function
- Example: **bar(** *variable,*
  col=*color,*
  border=*NA*,
  xlim=*range*,
  main=*title*,
  xlab=*x-axis,*
  ylab=*y-axis name***)**

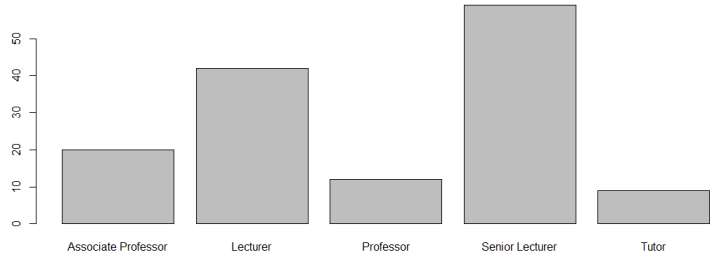




Presenting Data using barplot()

- To plot the academic staff data into a bar plot, use following command:

```
barplot(staffPosition.freq)
```



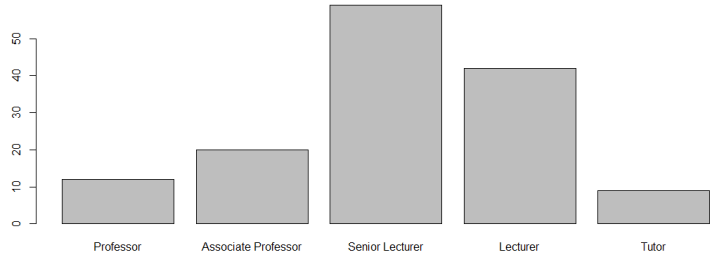
- What if we want to reorder it?
- Using following command in R we can reorder the data:

```
staffPosition = staffPosition[c(3,1,4,2,5)]
```

Reorder based on the index

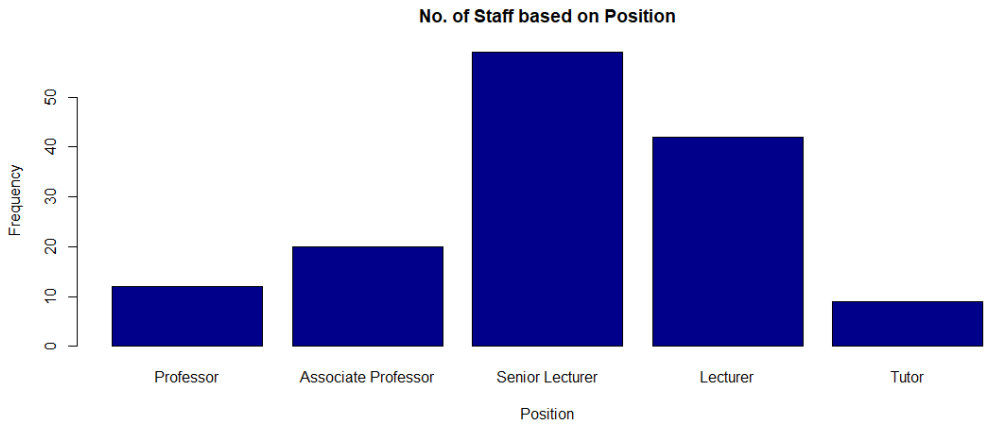
- Then replot again for the new ordered data:

```
barplot(staffPosition)
```



Settings in barplot()

- Settings of bar plot available in R include:
  - Color
  - Bar orientation
  - Range
  - Title, labels
- For example:




```
barplot(staffPosition, col="darkblue",
xlab="Position", ylab="Frequency", main="No. of
Staff based on Position")
```

Presenting Data using pie()

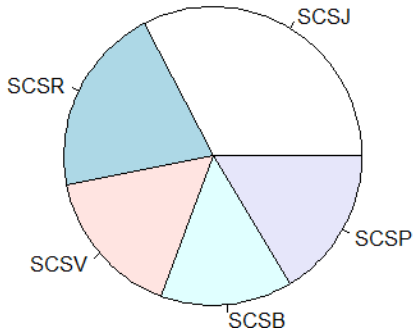
- Example: let say we have a set of data of number of students for 2017/2018 Intake as follow:
- Let say we import the data and save it into a variable name **Data2**
- After import the data in R, a pie chart can be plotted using following command:

Course	No. of students
SCSJ	80
SCSR	50
SCSV	40
SCSB	35
SCSP	40

```
pie(Data2$X2, Data2$X1)
```

Data to be plotted

Labels

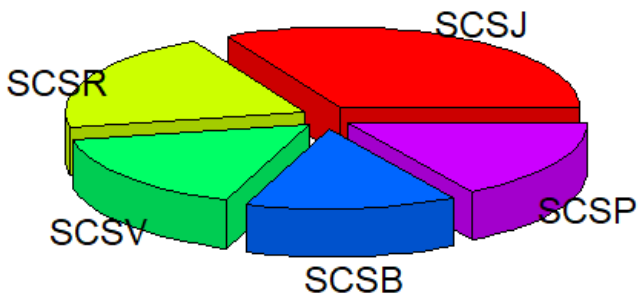


More with pie chart in R

- 3D pie chart is available with the use of package ***plotrix***
- After install the package, use the function **pie3D()**
- Example using the student data:

```
pie3D(Data2$X2, labels=Data2$X1,
      main="No. of Students For 2017/2018 Intake",
      explode=0.1)
```

"explode" parameter for the separation between each slice



Relative Frequency Distribution

- Refer back the academic staff data in previous slides.
- Let say we already have the frequency distribution in the variable ***staffPosition.freq***
- We can calculate the relative frequency distribution simply with following command

```
staffPosition.rfd <- staffPosition.freq/140
```

Position	Number of Staff (frequency)
Professor	12
Associate Professor	20
Senior Lecturer	59
Lecturer	40
Tutor	9
Total	140

- In R, arithmetic executed for a list variable will apply to each of the element.

```
> staffPosition.rfd <- staffPosition.freq/140
> staffPosition.rfd

         Professor Associate Professor     Senior Lecturer            Lecturer               Tutor
        0.08571429          0.14285714          0.42142857          0.28571429          0.06428571
> |
```

- Then plot the relative frequency distribution into a bar plot

Presenting Data using stem and leaf

- Let say we have a list of numbers in a variable named **num**:

  12, 13, 21, 27, 33, 34, 35, 37, 40, 40, 41

- In R in order to draw a stem-and-leaf plot, we can use the stem() function.

  stem(num)

```
> num <- c(12, 13, 21, 27, 33, 34, 35, 37, 40, 40, 41)
> stem(num)

  The decimal point is 1 digit(s) to the right of the |

  1 | 23
  2 | 17
  3 | 3457
  4 | 001
```

x: Please specify the data on which you want to draw the Stem and Leaf Plot. Here, you have to use the numeric vector, or a list containing numeric vector.

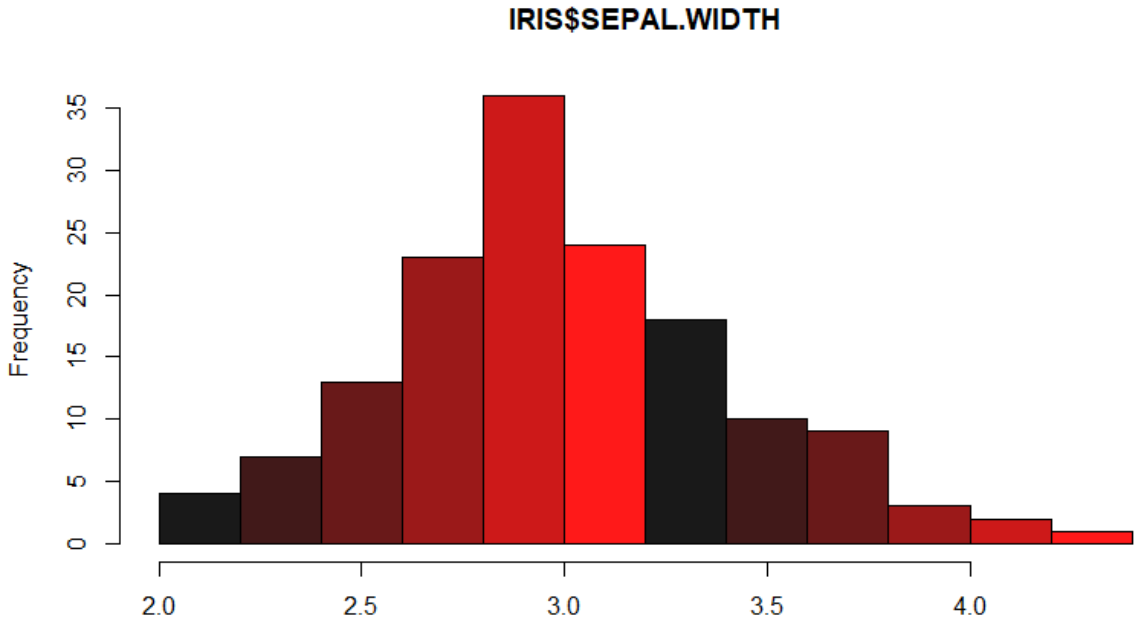
scale: Please specify the scale you want to use for your plot.

width: It is optional but you can use this to specify the desired width of a plot. By default it is 80

atom: It is a tolerance

Histogram in R

- In R, histogram is plotted using **hist()** function.
  - Example: **hist(***variable,* **col=***color,* **border=***NA*, **xlim=***range*, **main=***title*, **xlab=***x-axis*, **ylab***=y-axis name***)**



Presenting data in histogram

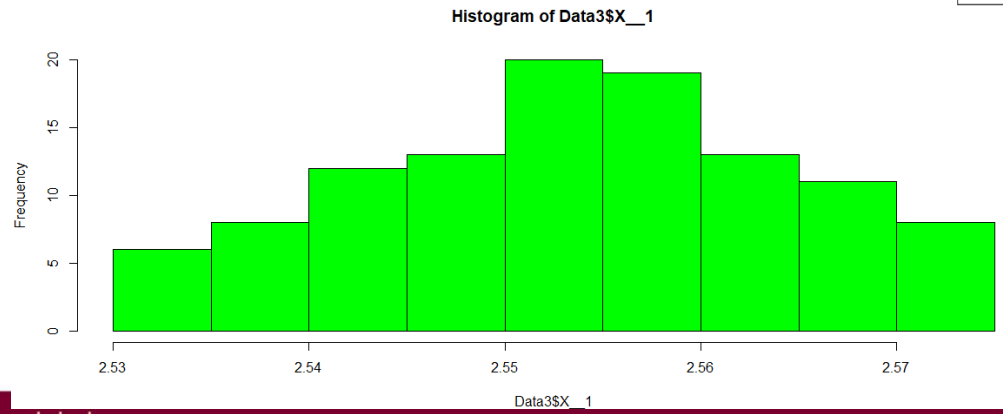
- Let say we have a group of data and imported into R in a variable named **Data3**
- To plot a histogram out of the data, we can use following:

```
hist(Data3$X1, col="green")
```

- We can save the output by assigning the command above into a variable.

```
output_h = hist(Data3$X1, col="green");
```

2.559	2.556	2.566	2.546	2.561	2.570	2.546
2.565	2.543	2.538	2.560	2.560	2.545	2.551
2.568	2.546	2.555	2.551	2.554	2.574	2.568
2.572	2.550	2.556	2.551	2.561	2.560	2.564
2.567	2.560	2.551	2.562	2.542	2.549	2.561
2.556	2.550	2.561	2.558	2.556	2.559	2.557
2.532	2.575	2.551	2.550	2.559	2.565	2.552
2.560	2.534	2.547	2.569	2.559	2.549	2.544
2.550	2.552	2.536	2.570	2.564	2.553	2.558
2.538	2.564	2.552	2.543	2.562	2.571	2.553
2.539	2.569	2.552	2.536	2.537	2.532	2.552
2.575 (h)	2.545	2.551	2.547	2.537	2.547	2.533
2.538	2.571	2.545	2.545	2.556	2.543	2.551
2.569	2.559	2.534	2.561	2.567	2.572	2.558
2.542	2.574	2.570	2.542	2.552	2.551	2.553
2.546	2.531 (l)	2.563	2.554	2.544		



Presenting data in histogram

- This variable holds the information calculated by R during the construction of the histogram.

```
> hist(Data3$X__1, col="green")
> output_h = hist(Data3$X__1, col="green")
> output_h
$breaks
 [1] 2.530 2.535 2.540 2.545 2.550 2.555 2.560 2.565 2.570 2.575

$counts
[1]  6  8 12 13 20 19 13 11  8

$density
[1] 10.90909 14.54545 21.81818 23.63636 36.36364 34.54545 23.63636 20.00000 14.54545

$mids
[1] 2.5325 2.5375 2.5425 2.5475 2.5525 2.5575 2.5625 2.5675 2.5725

$xname
[1] "Data3$X__1"

$equidist
[1] TRUE

attr(,"class")
[1] "histogram"
>
```

Interval calculated by R

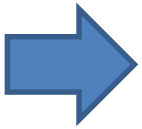
Grouped frequency

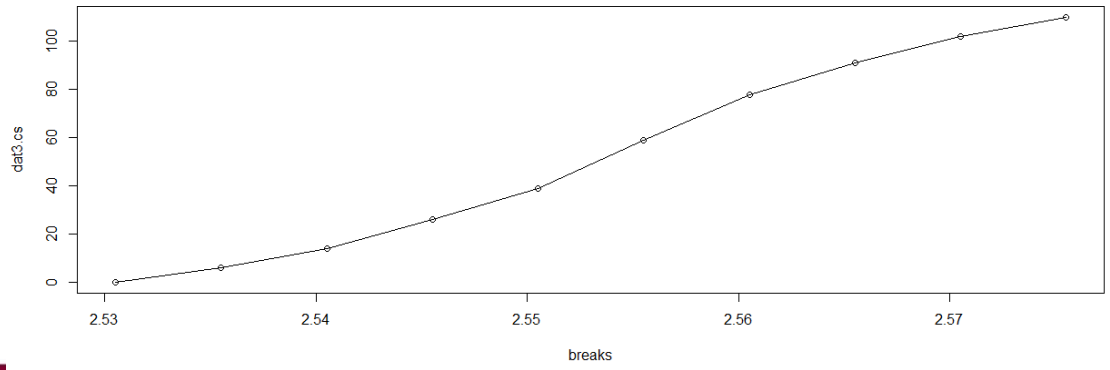
Mid points

Frequency distribution of grouped data

- Consider the data used in histogram plotting.
- We can group the data based on desired breaks.
- Generally follow steps below
  1. Extract the data (if it is in dataframe) into a list or vector
  2. Define the breaks (a.k.a interval)
  3. Group the data based on the breaks
  4. Calculated cumulative frequency
  5. Plot the result.

2.559	2.556	2.566	2.546	2.561	2.570	2.546
2.565	2.543	2.538	2.560	2.560	2.545	2.551
2.568	2.546	2.555	2.551	2.554	2.574	2.568
2.572	2.550	2.556	2.551	2.561	2.560	2.564
2.567	2.560	2.551	2.562	2.542	2.549	2.561
2.556	2.550	2.561	2.558	2.556	2.559	2.557
2.532	2.575	2.551	2.550	2.559	2.565	2.552
2.560	2.534	2.547	2.569	2.559	2.549	2.544
2.550	2.552	2.536	2.570	2.564	2.553	2.558
2.538	2.564	2.552	2.543	2.562	2.571	2.553
2.539	2.569	2.552	2.536	2.537	2.532	2.552
2.575 (h)	2.545	2.551	2.547	2.537	2.547	2.533
2.538	2.571	2.545	2.545	2.556	2.543	2.551
2.569	2.559	2.534	2.561	2.567	2.572	2.558
2.542	2.574	2.570	2.542	2.552	2.551	2.553
2.546	2.531 (l)	2.563	2.554	2.544		





Frequency distribution of grouped data

- STEP 1: Extract the data (if it is in dataframe) into a list or vector

```
Dat3 <- Data3$X1
```

Extract the info and assign into Dat3 which is a vector

- STEP 2: Define the breaks (a.k.a interval)

```
breaks <- seq(2.5305, 2.5375, 0.005)
```

(low limit, up limit, interval)

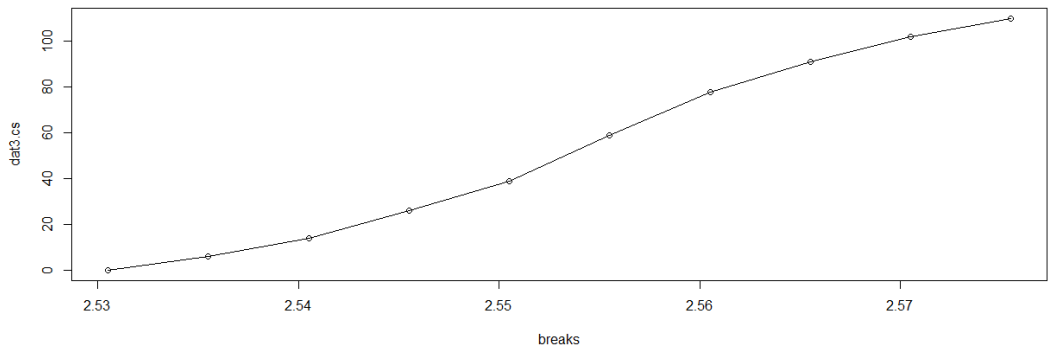
- STEP 3: Group the data based on the breaks

```
Dat3.cut <- cut(Dat3, breaks)
Dat3.freq <- table(Dat3.cut)
```

Categorize each data based on the breaks

Group the similar breaks and get the frequency

2.559	2.556	2.566	2.546	2.561	2.570	2.546
2.565	2.543	2.538	2.560	2.560	2.545	2.551
2.568	2.546	2.555	2.551	2.554	2.574	2.568
2.572	2.550	2.556	2.551	2.561	2.560	2.564
2.567	2.560	2.551	2.562	2.542	2.549	2.561
2.556	2.550	2.561	2.558	2.556	2.559	2.557
2.532	2.575	2.551	2.550	2.559	2.565	2.552
2.560	2.534	2.547	2.569	2.559	2.549	2.544
2.550	2.552	2.536	2.570	2.564	2.553	2.558
2.538	2.564	2.552	2.543	2.562	2.571	2.553
2.539	2.569	2.552	2.536	2.537	2.532	2.552
2.575 (h)	2.545	2.551	2.547	2.537	2.547	2.533
2.538	2.571	2.545	2.545	2.556	2.543	2.551
2.569	2.559	2.534	2.561	2.567	2.572	2.558
2.542	2.574	2.570	2.542	2.552	2.551	2.553
2.546	2.531 (l)	2.563	2.554	2.544		



Frequency distribution of grouped data

- STEP 4: Calculated cumulative frequency using **cumsum()**

```
Dat3.cs <- c(0, cumsum(Dat3.freq))
```

Calculate cumulative freq

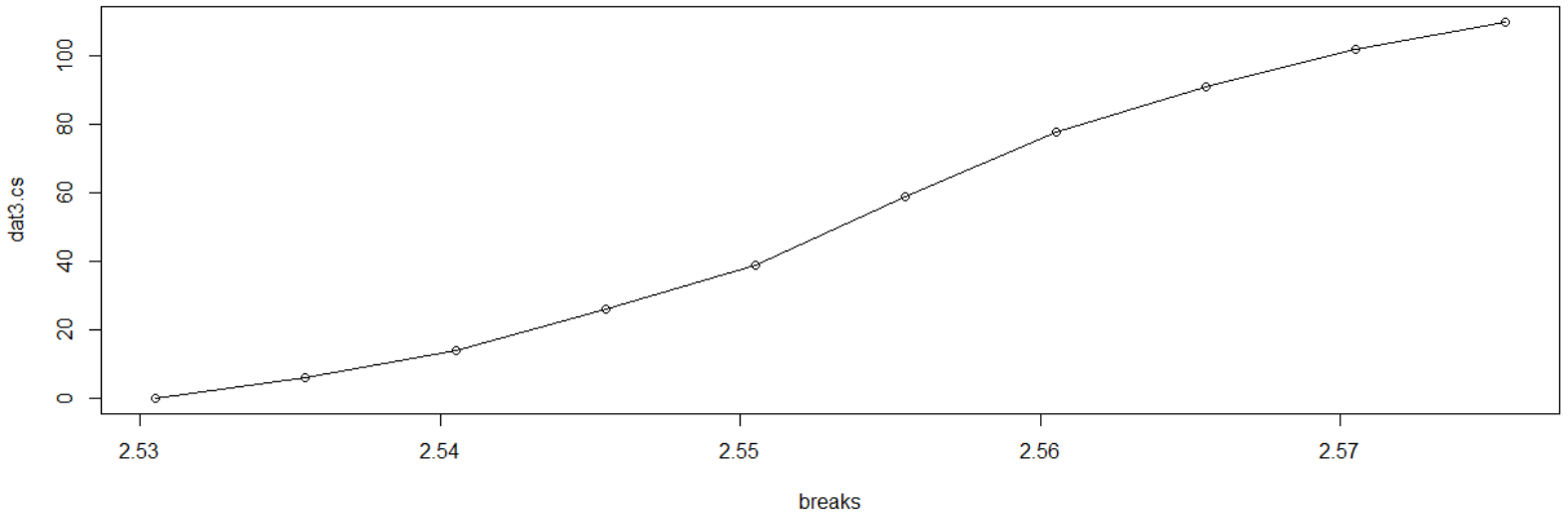
- STEP 5: Plot the results using **plot()** function

```
plot(breaks, Dat3.cs)
```

Plot a graph

```
lines(breaks, Dat3.cs)
```

Link the dots with line



Presenting Data in boxplot()

- Let say we have a set of data:

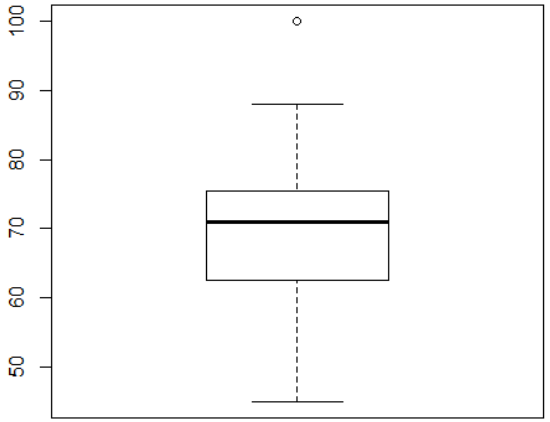
45 48 50 54 57 60 60 62 63 63 64 65 67 68 69
71 71 72 72 74 74 75 75 76 80 83 84 88 100 100
72 100

- First we have to make sure the data we have is in data frame in R. Let say we save this in a variable named **num2**
- Then to plot a boxplot, simply use following command:

```
boxplot(num2)
```

- We can also save the output of the boxplot function into a variable, let say **output**

```
output = boxplot(num2)
```



More in boxplot in R

- By assigning the output of a boxplot into a variable, information calculated by R can be revealed.

A list is returned which has

stats - having the position of the upper/lower extremes of the whiskers and box along with the median

n - the number of observation the boxplot is drawn with (notice that NA's are not taken into account)

conf - upper/lower extremes of the notch, out- value of the outliers

group - a vector of the same length as out whose elements indicate to which group the outlier belongs and

names - a vector of names for the groups.

```
> output <- boxplot(num2)
> output
$stats
     [,1]
[1,] 45.0
[2,] 62.5
[3,] 71.0
[4,] 75.5
[5,] 88.0

$n
[1] 31

$conf
         [,1]
[1,] 67.31091
[2,] 74.68909

$out
[1] 100 100 100

$group
[1] 1 1 1

$names
[1] "bdata"

>
```

Descriptive Statistics (Mean)

- Mean can be calculated using **mean()** function.
- For simple data such as a sequence of numbers. Let say as following:

```
num3 = c(20, 30, 40, 50, 80, 100)
```

- The mean can be calculated simply with following command:

```
mean(num3)
```

```
> num3 = c(20, 30, 40, 50, 80, 100)
>
> mean(num3)
[1] 53.33333
>
```

Descriptive Statistics (Median)

- Median can be calculated using **median()** function.
- Refer back to the same data:

```
num3 = c(20, 30, 40, 50, 80, 100)
```

- The median can be calculated simply with following command:

```
median(num3)
```

```
> num3 = c(20, 30, 40, 50, 80, 100)
>
> median(num3)
[1] 45
>
```

Descriptive Statistics (Mode)

- R **does not** have a standard in-built function to calculate mode. So we create a user function to calculate mode of a data set in R.
- Assume we have a list of numbers as follows:

  num4 = c(20, 30, 40, 50, 20, 80, 100)
- Mode can be calculated using following steps:
- STEP 1: get the available unique number using **unique()**

  unique.value <- unique(num4)
- STEP 2: find occurrence of unique number using **match()**

  unique.match <- match(num4, unique.value)

Descriptive Statistics (Mode)

- STEP 3: find the frequency of each unique number using **tabulate()**

```
unique.freq <- tabulate(unique.match)
```

- STEP 4: find the index which has the highest frequency using **which.max()**

```
unique.max <- which.max(unique.freq)
```

- STEP 5: get the value of the mode by using the index identified in STEP 4

```
num4_mode <- unique.value[unique.max]
```

```
> unique.value <- unique(num4)
> unique.value
[1]  20  30  40  50  80 100
> unique.match <- match(num4, unique.value)
> unique.match
[1] 1 1 2 3 4 5 6
> unique.freq <- tabulate(unique.match)
> unique.freq
[1] 2 1 1 1 1 1
> unique.max <- which.max(unique.freq)
> unique.max
[1] 1
> num4_mode <- unique.value[unique.max]
> num4_mode
[1] 20
>
```

Descriptive Statistics (Quartile)

- In R, information for quartile can be obtained using quantile() function.
- Let say we have a list of data as follows:

```
num5 = c(12, 4, 6, 11, 9, 15, 20, 18, 25, 30)
```

- The quartile information can be obtained via following:

```
quantile(num5, type=2)
```

- There are several types of setting for the function that will generate different results

```
> quantile(num5, type=2)
  0%  25%  50%  75% 100%
 4.0  9.0 13.5 20.0 30.0
> |
```

  - Source: https://tolstoy.newcastle.edu.au/R/e17/help/att-1067/Quartiles_in_R.pdf

Descriptive Statistics (Quartile)

- In R, information for quartile can be obtained using quantile() function.
- Let say we have a list of data as follows:

```
num5 = c(12, 4, 6, 11, 9, 15, 20, 18, 25, 30)
```

- The quartile information can be obtained via following:

```
quantile(num5, type=2)
```

- There are several types of setting for the function that will generate different results

```
> quantile(num5, type=2)
  0%  25%  50%  75% 100%
 4.0  9.0 13.5 20.0 30.0
> |
```

  - Source: https://tolstoy.newcastle.edu.au/R/e17/help/att-1067/Quartiles_in_R.pdf

Descriptive Statistics (variance / std dev)

- In R, variance can be calculated using the **var()** function.
- Using the same data as previous slide, variance can be obtained using the following command:

```
num5_variance <- var(num5)
```

```
> num5_variance <- var(num5)
> num5
 [1] 12  4  6 11  9 15 20 18 25 30
> num5_variance
[1] 69.11111
> |
```

- While for standard deviation, can be obtained using **sd()** in R
- For example:

```
num5_sd <- sd(num5)
```

```
> num5_sd <- sd(num5)
> num5
 [1] 12  4  6 11  9 15 20 18 25 30
> num5_sd
[1] 8.313309
> |
```

Skewness in R

- In R, skewness can be measured using **skewness()** function in the package named “moments”
- Let say we have a set of data as follows:

```
time <- c(19.09, 19.55, 17.89, 17.73, 25.15, 27.27, 25.24, 21.05,
          21.65, 20.92, 22.61, 15.71, 22.04, 22.60, 24.25)
```

- The skewness of the data can be measured using following command:

```
skewness(time)
```

```
> time <- c(19.09, 19.55, 17.89, 17.73, 25.15, 27.27, 25.24,
21.05,  21.65, 20.92, 22.61, 15.71, 22.04, 22.60, 24.25)
>
> skewness(time)
[1] -0.01565162
>
```

Kurtosis in R

- In R, kurtosis can be measured using **kurtosis()** function in the package named "moments" as same as previous.
- Let say we use the same data as previous slide:

```
time <- c(19.09, 19.55, 17.89, 17.73, 25.15, 27.27, 25.24, 21.05,
          21.65, 20.92, 22.61, 15.71, 22.04, 22.60, 24.25)
```

- The kurtosis of the data can be measured using following command:

```
kurtosis(time)
```

```
> time <- c(19.09, 19.55, 17.89, 17.73, 25.15, 27.27, 25.24,
21.05,  21.65, 20.92, 22.61, 15.71, 22.04, 22.60, 24.25)
>
> kurtosis(time)
[1] 2.301051
>
```

R Programming Tutorial (Part 2)

Prepared by: Chan Weng Howe

Outline

- Interval Estimate in R
- Hypothesis Testing in R

Interval Estimate in R for

Example

- A recent study investigated the effects of a "Buckle Up Your Toddlers" campaign to get parents to use the grocery cart seat belts. Investigators observed a representative sample of parents at grocery stores in a large city, and found 192 out of 594 parents buckling up their toddlers.

a. Compute a point estimate of the true proportion of all parents who buckle up their toddlers.

b. Construct and interpret a 95% confidence interval for, the true proportion of all parents who buckle up their toddlers.

```
*(1-pbar)/n)
75)
*cv;

erval
gin,
```

```
0.3608448
```

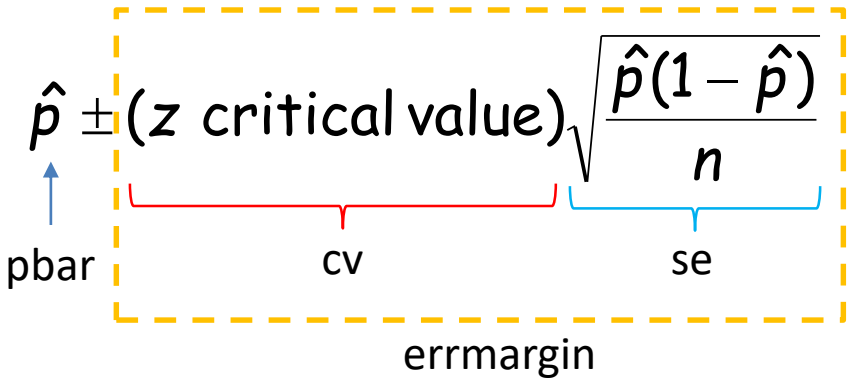
Interval Estimate in R for Proportions

WHAT WE KNOW?

- Sample size, n: 594
- 192 of the sample buckling up their toddlers.

WHAT WE WANT TO FIND?

- Point estimate of the true proportion.
- 95% confidence interval for the true proportion.



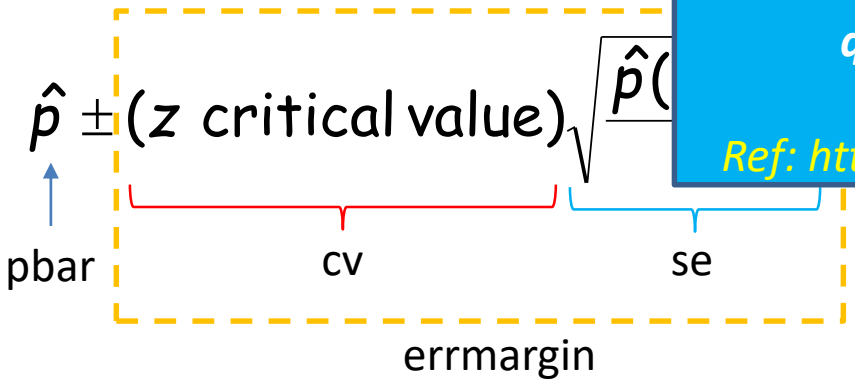

```
1         n = 594
2         k = 192
3         pbar = n/k;
4         se = sqrt(pbar*(1-pbar)/n)
5         cv = qnorm(0.975)
6         errmargin = se*cv;
8
9         // Get the interval
10        pbar+c(-errmargin,
errmargin)
11
12
13
14
15
16
17
18
19
20
21
22
```

```
[1] 0.2856198 0.3608448
```

Interval Estimate in R for Proportions

$1-\alpha$	α	$\alpha/2$	$z_{\alpha/2}$
.90	.10	.05	$z_{.05} = 1.645$
.95	.05	.025	$z_{.025} = 1.96$
.98	.02	.01	$z_{.01} = 2.33$
.99	.01	.005	$z_{.005} = 2.575$

true proportion.




```
n = 594
k = 192
pbar = n/k;
se = sqrt(pbar*(1-pbar)/n)
cv = qnorm(0.975)
errmargin = se*cv;

// Get the interval
```

sed to get the Z critical value. E.g.: for 95% confidence level, *a/2* = 0.025, however, parameter to put in ***qnorm()*** is 1-0.025 = 0.975

Ref: http://seankross.com/notes/dpqr/

```
18
19
20
21
22
```

```
[1] 0.2856198 0.3608448
```

Interval Estimate in R for Proportions

WHAT WE KNOW?

- Sample size, n: 594
- 192 of the sample bu
  toddlers.

WHAT WE WANT TO FIN

- Point estimate of the
  proportion.
- 95% confidence inter
  true proportion.

$\hat{p} \pm (z \text{ critical valu}$

pbar cv

errm

```
1          n = 594
```

UTM UNIVERSITI TEKNOLOGI MALAYSIA

www.utm.my

Solution

a) point estimate: p = 192 / 594 = 0.3232

b) 95% CI: = $95\%\ \text{CI: } 0.3232 \pm 1.960\sqrt{\frac{0.3232(0.6768)}{594}} = (0.2856, 0.3608).$

I am 95% confident that the true proportion of parents who buckle up their toddlers is between 0.286 and 0.361.

Interval Estimate in R for Proportions (Alternative)

prop.test() is used to test proportions.
It includes the calculation of confidence interval (95% by default)

How to use *prop.test()*:

```
prop.test(x, n, p = NULL,
          alternative = c("two.sided", "less", "greater"),
          conf.level = 0.95, correct = TRUE)
```

Ref: https://stat.ethz.ch/R-manual/R-devel/library/stats/html/prop.test.html

```
1          n = 594
2          k = 192
3
4          prop.test(k, n)
5
6
7
8
9
10
```

```
        1-sample proportions test with continuity
correction

data:  k out of n, null probability 0.5
X-squared = 73.537, df = 1, p-value < 2.2e-16
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.2860486 0.3627384
sample estimates:
        p
0.3232323
```

Confidence Interval (Default: 95%)

Interval Estimate in R for Mean

WHAT WE KNOW?

```
n = 25
75
, 374, 309, 499,
, 361, 514, 462,
, 439, 348, 344,
, 374, 302, 466,
, 316, 296, 332,
an(x)
ev/sqrt(n)
m(0.975)
 = cv*se
I.
rrmargin,
```

```
7605 399.5595
```

UTM UNIVERSITI TEKNOLOGI MALAYSIA

Example

www.utm.my

- A computer company samples demand during lead time over 25 time periods:

235	374	309	499	253
421	361	514	462	369
394	439	348	344	330
261	374	302	466	535
386	316	296	332	334

Its is known that the standard deviation of demand over lead time is 75 computers. We want to estimate the ***mean*** demand over lead time with 95% confidence in order to set inventory levels

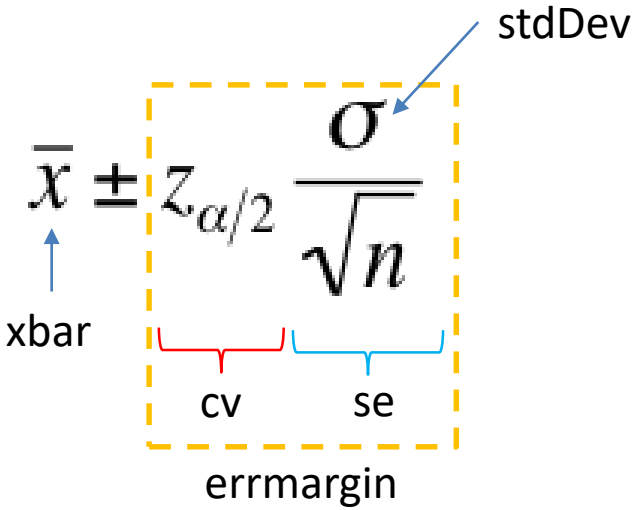
Interval Estimate in R for Mean

WHAT WE KNOW?

- Sample size, n: 25
- Standard deviation: 75

WHAT WE WANT TO FIND?

- 95% confidence interval for the mean.




```
1         n = 25
2         stdDev = 75
3         x = c(235, 374, 309, 499,
253,
4               421, 361, 514, 462,
369,
5               394, 439, 348, 344,
330,
6               261, 374, 302, 466,
535,
7               386, 316, 296, 332,
334)
8         xbar = mean(x)
9         se = stdDev/sqrt(n)
10        cv = qnorm(0.975)
11        errmargin = cv*se
12
13        // 95% C.I.
14        xbar+c(-errmargin,
errmargin)
15
16
17
18
19
20
```

```
[1] 340.7605 399.5595
```

Interval Estimate in R for Mean

WHAT WE KNOW?

- Sample size, n: 25
- Standard deviation

WHAT WE WANT TO F

- 95% confidence int
  mean.

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

xbar

cv

s

errmargi

Solution

www.utm.my

- In order to use our confidence interval estimator, we need the following pieces of data:

$\bar{x}$	370.16
$z_{\alpha/2}$	1.96
σ	75
n	25

Calculated from the data…

$1-\alpha = .95,\ \therefore \alpha/2 = .025$
$so\ z_{\alpha/2} = z_{.025} = 1.96$

Given

therefore: $\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = 370.16 \pm z_{.025}\frac{75}{\sqrt{25}} = 370.16 \pm 1.96\frac{75}{\sqrt{25}} = 370.16 \pm 29.40$

The **lower** and **upper** confidence limits are 340.76 and 399.56.

innovative • entrepreneurial • global 34

19
20

Hypothesis Testing

Decision Criterion

- Traditional method:
  - If the test statistic falls within the critical region, reject H_0.
  - If the test statistic does not fall within the critical region, fail to reject H_0.

Hypothesis Testing

METHOD 1

1. Identify critical region
2. Calculate test statistics
3. If test statistics fall within critical region, reject H_0
4. If test statistics does not fall within critical region, fail to reject H_0

METHOD 2

1. Define significance level.
2. Calculate test statistics.
3. Find p-value
4. If p-value < significance level, reject H_0
5. If p-value > significance level, fail to reject H_0

Hypothesis Testing in R

UTM UNIVERSITI TEKNOLOGI MALAYSIA

www.utm.my

Example

- **Writing a Hit Song** In the manual "How to Have a Number One the Easy Way," by KLF Publications, it is stated that a song "must be no longer than three minutes and thirty seconds" (or 210 seconds). A simple random sample of 40 current hit songs results in a mean length of 252.5 sec. (The songs are by Timberlake, Furtado, Daughtry, Stefani, Fergie, Akon, Ludacris, etc.) Assume that the standard deviation of song lengths is 54.5 sec. Use a 0.05 significance level to test the claim that the sample is from a population of songs with a mean greater than 210 sec. What do these results suggest about the advice given in the manual?


```
z test

)/(sd/sqrt(n))

 critical value
orm(1-alpha)
```

Hypothesis Testing in R

Population mean with Known Variance – One Tailed (Traditional Method)

WHAT WE KNOW?

- Sample size, n: 40
- Standard deviation: 54.5
- Sample mean: 252.5
- H_0: μ = 210 H_1: μ > 210
- Significance level: 0.05

$$z = \frac{x - \mu_x}{\frac{\sigma}{\sqrt{n}}}$$

```
1         n = 40
2         sd = 54.5
3         xbar = 252.5
4         mu = 210
5         alpha = 0.05
6
7         // Calculate z test
8statistics
9         z = (xbar-mu)/(sd/sqrt(n))
10
11        // Calculate critical value
12        z.alpha = qnorm(1-alpha)
13
14
15
16   > z
17   [1] 4.931993
18   > z.alpha
19   [1] 1.644854
20
21
22
```

Hypothesis Testing in R

Population mean with Known Variance – One Tailed (P-value Method)

WHAT WE KNOW?

- Sample size, n: 40
- Standard deviation: 54.5
- Sample mean: 252.5
- H_0: μ = 210 H_1: μ > 210
- Significance level: 0.05

$$z = \frac{x - \mu_x}{\frac{\sigma}{\sqrt{n}}}$$

```
1         n = 40
2         sd = 54.5
3         xbar = 252.5
4         mu = 210
5         alpha = 0.05
6
7         // Calculate z test
statistics
9         z = (xbar-mu)/(sd/sqrt(n))
10
11        // Calculate p-value of z
12        pval = pnorm(z,
19
20  [1] 4.069748e-07
21
22
```

pnorm() can be used to get p-value for a calculated test statistics.

Ref: http://seankross.com/notes/dpqr/

Hypothesis Testing in R

Population
Variance – C
(P-value Me

WHAT WE KN

- Sample siz
- Standard d
- Sample m
- H_0: μ = 21
- Significanc

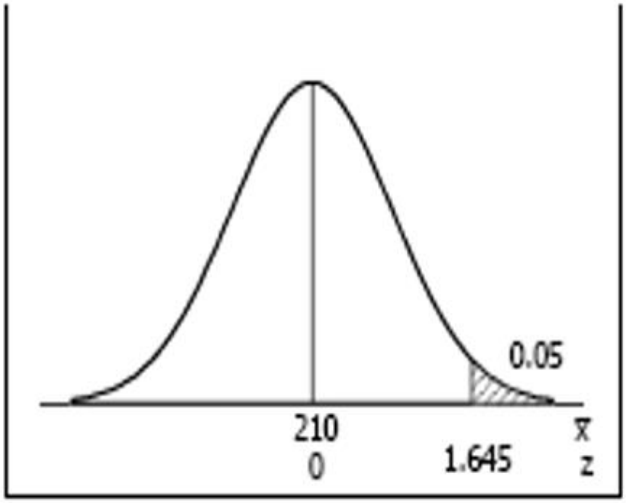
$z =$ _

Solution

original claim: $\mu > 210$ sec
H_o: $\mu = 210$ sec
H_1: $\mu > 210$ sec
$\alpha = 0.05$
C.V. $z = z_\alpha = z_{0.05} = 1.645$
calculations:

$$z_{\bar{x}} = (\bar{x} - \mu_{\bar{x}})/\sigma_{\bar{x}}$$
$$= (252.5 - 210)/(54.5/\sqrt{40})$$
$$= 42.5/8.6172 = 4.93$$

P-value = $P(z > 4.93) = 0.5 - 0.4999 = 0.0001$



conclusion:

Reject H_o; there is sufficient evidence to conclude that $\mu > 210$. There is sufficient evidence to support the claim that the sample is from a population of songs with a mean greater than 210 seconds. These results suggest that the advice given in the manual is not good advice.

innovative ● entrepreneurial ● global 51

Hypothesis Testing in R

Population mean with Known Variance – Two Tailed (Traditional Method)

Problem:

Suppose the mean weight of King Penguins found in an Antarctic colony last year was 15.4 kg. In a sample of 35 penguins same time this year in the same colony, the mean penguin weight is 14.6 kg. Assume the population standard deviation is 2.5 kg. At .05 significance level, can we reject the null hypothesis that the mean penguin weight does not differ from last year?

```
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
```

Hypothesis Testing in R

Population mean with Known Variance – Two Tailed (Traditional Method)

WHAT WE KNOW?

- Sample size, n: 35
- Standard deviation: 2.5
- Sample mean: 14.6
- H_0: μ = 15.4 H_1: μ ≠ 15.4
- Significance level: 0.05

$$z = \frac{x - \mu_x}{\frac{\sigma}{\sqrt{n}}}$$

```
n = 35
sd = 2.5
xbar = 14.6
mu = 15.4
alpha = 0.05

// Calculate Z statistics
z = (xbar-mu)/(sd/sqrt(n))

// Calculate critical value
z.alpha = qnorm(1-(alpha/2))

// Display the c.v. on both sides
c(-z.alpha, z.alpha)
```

```
> Z
[1] -1.893146
> c(-z.alpha, z.alpha)
[1] -1.959964  1.959964
```

Hypothesis Testing in R

Population mean with Known Variance – Two Tailed (Traditional Method)

WHAT WE KNOW?

- Sample size, n: 35
- Standard deviation: 2.5
- Sample mean: 14.6
- H_0: μ = 15.4 H_1: μ ≠ 15.4
- Significance level: 0.05

In two tailed test, the critical value is calculated based on half of the alpha value

$$\frac{\sigma}{\sqrt{n}}$$

```
1        n = 35
2        sd = 2.5
3        xbar = 14.6
4        mu = 15.4
5        alpha = 0.05
6
7        // Calculate Z statistics
8        z = (xbar-mu)/(sd/sqrt(n))
9
10       // Calculate critical value
11       z.alpha = qnorm(1-(alpha/2))
12
13       // Display the c.v. on both
14 sides
15       c(-z.alpha, z.alpha)
16
19
20
21
22
```

```
> Z
[1] -1.893146
> c(-z.alpha, z.alpha)
[1] -1.959964  1.959964
```

Hypothesis Testing in R

Population mean with Known Variance – Two Tailed (P-value Method)

WHAT WE KNOW?

- Sample size, n: 35
- Standard deviation: 2.5
- Sample mean: 14.6
- H_0: μ = 15.4 H_1: μ ≠ 15.4
- Significance level: 0.05

$$z = \frac{x - \mu_x}{\frac{\sigma}{\sqrt{n}}}$$

```
1        n = 35
2        sd = 2.5
3        xbar = 14.6
4        mu = 15.4
5        alpha = 0.05
6
7        // Calculate Z statistics
8        z = (xbar-mu)/(sd/sqrt(n))
9
10       // Calculate p-value for Z
11       pval = 2*pnorm(z)
12
13
14
15
16
17
18
19
20   [1] 0.05833852
21
22
```

Hypothesis Testing in R

Population mean with Known Variance – Two Tailed (P-value Method)

WHAT WE KNOW?

- Sample size, n: 35
- Standard deviation: 2.5
- Sample mean: 14.6
- H_0: μ = 15.4 H_1: μ ≠ 15.4
- Significance level: 0.05

P-value calculated by pnorm() represent one-tailed P-value. For two tailed test, the P-value must multiply by 2

$$\frac{}{\sqrt{n}}$$

```
1        n = 35
2        sd = 2.5
3        xbar = 14.6
4        mu = 15.4
5        alpha = 0.05
6
7        // Calculate Z statistics
8        z = (xbar-mu)/(sd/sqrt(n))
9
10       // Calculate p-value for Z
11       pval = 2*pnorm(z)
12
13
14
19
20   [1] 0.05833852
21
22
```

Hypothesis Testing in R

Population mean with Unknown Variance – One Tailed (Traditional Method)

WHAT WE KNOW?

- Sample size, n: 15
- List of data points given
- H_0: μ = 0.82 H_1: μ > 0.82
- Significance level: 0.05

```
1         x = c(0.8411, 0.8191,
2  0.8182,                0.8125,
3  0.8750, 0.8580,
4  0.8532, 0.8483, 0.8276,
5                0.7983, 0.8042,
6  0.8730,
7                0.8282, 0.8359,
8  0.8660)
9         n = 15
10        s = sd(x)
11        xbar = mean(x)
12        mu = 0.82
13
14        // Calculate t statistics
          t = (xbar-mu)/(s/sqrt(n))

          // Calculate critical value
          alpha = 0.05
```

qt() can be used to get the T critical value.

2 parameters: 1) t statistics, 2) df

Ref: http://stat.ethz.ch/R-manual/R-devel/library/stats/html/TDist.html

```
> t
[1] 2.718979
> t.alpha
[1] 1.76131
```

Hypothesis Testing in R

UTM UNIVERSITI TEKNOLOGI MALAYSIA

www.utm.my

Example

The increased availability of light materials with high strength has revolutionized the design and manufacture of golf clubs, particularly drivers. Clubs with hollow heads and very thin faces can result in much longer tee shots, especially for players of modest skills. This is due partly to the "spring-like effect" that the thin face imparts to the ball. Firing a golf ball at the head of the club and measuring the ratio of the outgoing velocity of the ball to the incoming velocity can quantify this spring-like effect. The ratio of velocities is called the coefficient of restitution of the club. An experiment was performed in which 15 drivers produced by a particular club maker were selected at random and their coefficients of restitution measured. In the experiment the golf balls were fired from an air cannon so that the incoming velocity and spin rate of the ball could be precisely controlled. It is of interest to determine if there is evidence (with $\alpha = 0.05$) to support a claim that the mean coefficient of restitution exceeds 0.82. The observations follow:

0.8411	0.8191	0.8182	0.8125	0.8750
0.8580	0.8532	0.8483	0.8276	0.7983
0.8042	0.8730	0.8282	0.8359	0.8660

innovative • entrepreneurial • global

57

```
x = c(0.8411, 0.8191,
      0.8125,

.8276,
7983, 0.8042,

3282, 0.8359,

ean(x)
2

late t statistics
r-mu)/(s/sqrt(n))

late critical value
.05

979

31
```

Hypothesis Testing in R

Population mean with Unknown Variance – One Tailed (Traditional Method)

WHAT WE KNOW?

- Sample size, n: 15
- List of data points given
- H_0: μ = 0.82 H_1: μ > 0.82
- Significance level: 0.05

$$T_0 = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$$

```
x = c(0.8411, 0.8191,
0.8182,                0.8125,
0.8750, 0.8580,
0.8532, 0.8483, 0.8276,
             0.7983, 0.8042,
0.8730,
             0.8282, 0.8359,
0.8660)
n = 15
s = sd(x)
xbar = mean(x)
mu = 0.82

// Calculate t statistics
t = (xbar-mu)/(s/sqrt(n))

// Calculate critical value
alpha = 0.05
```

```
> t
[1] 2.718979
> t.alpha
[1] 1.76131
```

Hypothesis Testing in R

Population Variance – (Traditiona

WHAT WE K

- Sample si
- List of dat
- H_0: μ = 0.
- Significan

$$T_0 =$$

UTM UNIVERSITI TEKNOLOGI MALAYSIA

www.utm.my

Solution

- Reject H_0 if $t_0 > t_{0.05,14} = 1.761$.
- Computations: Since = 0.83725, s = 0.02456, $\mu_0 = 0.82$ and $n = 15$, we have

$$t_0 = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{0.83725 - 0.82}{0.02456/\sqrt{15}} = 2.72.$$

- Conclusion: Since $t_0 = 2.72 > 1.761$, we reject H_0: μ = 0.82 at the 0.05 level of significance that the mean coefficient of restitution exceeds 0.82.

Hypothesis Testing in R

Population mean with Unknown Variance – One Tailed (P-value Method)

WHAT WE KNOW?

- Sample size, n: 15
- List of data points given
- H_0: μ = 0.82 H_1: μ > 0.82
- Significance level: 0.05

```
x = c(0.8411, 0.8191, 0.8182, 0.8125, 0.8750, 0.8580, 0.8532, 0.8483, 0.8276, 0.7983, 0.8042, 0.8730, 0.8282, 0.8359, 0.8660)
n = 15
s = sd(x)
xbar = mean(x)
mu = 0.82

// Calculate t statistics
t = (xbar-mu)/(s/sqrt(n))

// Calculate p-value of t
pval = pt(t, df=n-1,
```

```
[1] 0.008313369
```

pt() is used to get the p-value of T statistics.

2-3 parameters: 1) t statistics, 2) df 3) lower.tail settings

Ref: http://stat.ethz.ch/R-manual/R-devel/library/stats/html/TDist.html

Hypothesis Testing in R

Population proportion – One Tailed (Traditional Method)

PROBLEM

Suppose 60% of citizens voted in last election. 85 out of 148 people in a telephone survey said that they voted in current election. At 0.5 significance level, can we reject the null hypothesis that the proportion of voters in the population is below 60% this year?

$$z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}}$$

```
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
```

Hypothesis Testing in R

Population proportion –
One Tailed (Traditional Method)

WHAT WE KNOW?

- Sample size, n: 148
- People who vote : 85
- H_0: p = 0.6 H_1: p < 0.6
- Significance level: 0.05

$$z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}}$$

```
1        n = 148
2        k = 85
3        p = 0.6
4        pbar = k/n
5
6        // Calculate Z statistics
7        z = (pbar-p)/sqrt(p*(1-p)/n)
8
9        // Calculate critical value
10       alpha = 0.05
11       z.alpha = qnorm(1-alpha)
12       -z.alpha
13
14
15
16   > z
17   [1] -0.6375983
18   > -z.alpha
19
20   [1] -1.644854
21
22
```

Hypothesis Testing in R

Population proportion –
One Tailed (P-value Method)

WHAT WE KNOW?

- Sample size, n: 148
- People who vote : 85
- H_0: p = 0.6 H_1: p < 0.6
- Significance level: 0.05

$$z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}}$$

```
1        n = 148
2        k = 85
3        p = 0.6
4        pbar = k/n
5
6        // Calculate Z statistics
7        z = (pbar-p)/sqrt(p*(1-p)/n)
8
9        // Calculate p-value for Z
10       pval = pnorm(z)
11
12
13
14
15
16
17
18
19
20
21       [1] 0.2618676
22
```

Hypothesis Testing in R

Population proportion – Two Tailed

PROBLEM

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

$$z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}}$$

```
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
```

Hypothesis Testing in R

Population variance – Two Tailed

WHAT WE KNOW?

- Sample size, n: 9
- Sample standard deviation: 3
- $H_0: \sigma = 3.7$ $H_1: \sigma < 3.7$

qchisq() is used to get calculate chi-square statistics

2 parameters: 1) significance level, 2) df

Ref:http://personal.psu.edu/drh20/astrostatistics/excerpts/html/stats/html/Chisquare.html

```
1         n = 9
2         s.sd = 3
3         sigma = 3.7
4         alpha = 0.05
5
6         // Calculate chi-square
7         x2 = ((n-
8 1)*s.sd^2)/(sigma^2)
9
10        // Calculate critical value
          x2.alpha = qchisq(alpha,
  =n-1)
21
22
```

```
> x2
[1] 5.259313
> x2.alpha
[1] 2.732637
```

Hypothesis Testing in R

Population variance – Two Tailed

WHAT WE KNOW?

- Sample size, n: 9
- Sample standard deviation: 3
- H_0: $\sigma = 3.7$ H_1: $\sigma < 3.7$
- Significance level: 0.05

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

```
1         n = 9
2         s.sd = 3
3         sigma = 3.7
4         alpha = 0.05
5
6         // Calculate chi-square
7         x2 = ((n-
8 1)*s.sd^2)/(sigma^2)
9
10        // Calculate critical value
11        x2.alpha = qchisq(alpha,
12 df=n-1)
13
14
15
16  > x2
17  [1] 5.259313
18  > x2.alpha
19  [1] 2.732637
20
21
22
```

References

- http://www.r-tutor.com/

THE END…

R Programming Tutorial (Part 3)

Prepared by: Chan Weng Howe

Outline

- Hypothesis Testing in R (2 samples)
  - Two proportions
  - Two means: independent samples
    - Variance / Standard deviations Known
    - Variance / Standard deviations Unknown
      - Equal Variance
      - Unequal Variance

Two Proportions

UTM UNIVERSITI TEKNOLOGI MALAYSIA

Example

- Below are the sample data and summary statistics:

	Group 1	**Group 2**
	Subjects Given $1 Bill	Subjects Given 4 Quarters
Spent the money	$x_1=12$	$x_2=27$
Subjects in group	$n_1=46$	$n_2=43$

$$\hat{p}_1 = \frac{12}{46} \qquad \hat{p}_2 = \frac{27}{43}$$

$$\bar{p} = \frac{12+27}{46+43} = 0.438202$$

+n2)

/n1)+(pbar*qbar

pha)

.64485362695147"
8739527429506"

Two Proportions

WHAT WE KNOW?

- Two samples:
  - x_1=12, x_2=27
  - n_1=46, n_2=43
- Significance level: 0.05

WHAT WE WANT TO FIND?

$$H_0 : p_1 = p_2$$

$$H_1 : p_1 < p_2$$

$$\bar{p} = \frac{x_1 + x_2}{n_1 + n_2} \qquad \bar{q} = 1 - \bar{p}$$

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{\bar{p}\bar{q}}{n_1} + \frac{\bar{p}\bar{q}}{n_2}}}$$

```
1    x1 = 12
2    x2 = 27
3    n1 = 46
4    n2 = 43
5
6    phat1 <- x1/n1
7    phat2 <- x2/n2
8
9    pbar = (x1+x2)/(n1+n2)
10   qbar = 1-pbar
11
12   z = ((phat1-phat2)-
13   0)/sqrt((pbar*qbar/n1)+(pbar*qbar
14   /n2))
15
16   alpha = 0.05
17   z.alpha = qnorm(alpha)
18
19
20
21   [1] "Critical Value: -1.64485362695147"
22   [1] "Z statistics: -3.48739527429506"
```

Two Proportions

UTM UNIVERSITI TEKNOLOGI MALAYSIA

www.utm.my

Exercise

- Time magazine reported the result of a telephone poll of 800 adult Americans. The question posed of the Americans who were surveyed was: "Should the federal tax on cigarettes be raised to pay for health care reform?" The results of the survey were:

Non-smoker	Smoker
N_1=605	**N_2=195**
Y1=351 said yes	**Y2=41 said yes**

- Is there sufficient evidence at the α = 0.05 level, say, to conclude that the two populations — smokers and non-smokers — differ significantly with respect to their opinions?

```
+n2)

-
/n1)+(pbar*qbar

pha/2)

5996398454005"
22953986016"
```

Two Proportions

WHAT WE KNOW?

- Two samples:
  - x_1=351, x_2=41
  - n_1=605, n_2=195
- Significance level: 0.05

WHAT WE WANT TO FIND?

$$H_0 : p_1 = p_2$$

$$H_1 : p_1 \neq p_2$$

$$\bar{p} = \frac{x_1 + x_2}{n_1 + n_2} \qquad \bar{q} = 1 - \bar{p}$$

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{\bar{p}\bar{q}}{n_1} + \frac{\bar{p}\bar{q}}{n_2}}}$$

```
1    x1 = 351
2    x2 = 195
3    n1 = 605
4    n2 = 195
5
6    phat1 <- x1/n1
7    phat2 <- x2/n2
8
9    pbar = (x1+x2)/(n1+n2)
10   qbar = 1-pbar
11
12   z = ((phat1-phat2)-
13   0)/sqrt((pbar*qbar/n1)+(pbar*qbar
14   /n2))
15
16   alpha = 0.05
17   z.alpha = qnorm(alpha/2)
18
19
20
21   [1] "Critical Value: -1.95996398454005"
22   [1] "Z statistics: -10.9522953986016"
```

Two Means, independent samples, known σ^2

WHAT WE KNOW?

UTM UNIVERSITI TEKNOLOGI MALAYSIA

www.utm.my

Example

A product developer is interested in reducing the drying time of a primer paint. Two formulations of the paint are tested; formulation 1 is the standard chemistry, and formulation 2 has a new drying ingredient that should reduce the drying time. From experience, it is known that the standard deviation of drying time is 8 minutes, and this inherent variability should be unaffected by the addition of the new ingredient. Ten specimens are painted with formulation 1, and another ten specimens are painted with formulation 2; the 20 specimens are painted in random order and normally distributed. The two sample average drying times are $\bar{x}_1$ = 121 minutes and $\bar{x}_2$ = 112 minutes, respectively. What conclusions can the product developer draw about the effectiveness of the new ingredient, using $\alpha = 0.05$?

INSPIRING CREATIVE AND INNOVATIVE MINDS


```
gma2^2
gma2^2

ar2-
ance1/n1)+(variance

m(alpha,
SE)

e: 1.64485362695147"
 2.51557647468726"
```

Two Means, independent samples, known σ^2

WHAT WE KNOW?

- Two samples:
  - $\sigma_1 = \sigma_2 = 8$
  - $n_1=10$, $n_2=10$
  - $\bar{x}_1 = 121$, $\bar{x}_2 = 112$

WHAT WE WANT TO FIND?

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

$$z_0 = \frac{\bar{x}_1 - \bar{x}_2 - 0}{\sqrt{\frac{\sigma_1^2}{n1} + \frac{\sigma_2^2}{n2}}}$$

```
1    xbar1 = 121
2    xbar2 = 112
3
4    n1 = 10
5    n2 = 10
6
7    sigma1 = 8
8    sigma2 = 8
9
10   variance1 = sigma2^2
11   variance2 = sigma2^2
12
13   z = ((xbar1-xbar2-
14   0)/(sqrt((variance1/n1)+(variance
15   2/n2))))
16
17   alpha = 0.05
18   z.alpha = qnorm(alpha,
19   lower.tail=FALSE)
20
21   [1] "Critical Value: 1.64485362695147"
22   [1] "Z statistics: 2.51557647468726"
```

Two Means, independent samples, unknown σ^2 (Case: equal σ^2)

WHAT WE KNOW?

Case 1 : $\sigma_1^2 = \sigma_2^2 = \sigma^2$

- **Example**

Two catalysts are being analyzed to determine how they affect the mean yield of a chemical process. Specifically, catalyst 1 is currently in use, but catalyst 2 is acceptable. Since catalyst 2 is cheaper, it should be adopted, providing it does not change the process yield. A test is run in the pilot plant and results in the data shown in Table 5.1. Is there any difference between the mean yields? Use α = 0.05, and assume equal variances.


```
sigma2^2
sigma2^2

-xbar2-
ariance1/n1)+(variance

5
norm(alpha,
FALSE)
```

```
.1447866879178 2.1447866879178"
5359086434619"
```

Two Means, independent samples, unknown σ^2 (Case: equal σ^2)

WHAT WE KNOW?

- Two samples:
  - Two sets of data
  - Significance: 0.05

WHAT WE WANT TO FIND?

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

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

$$t_0 = \frac{\bar{x}_1 - \bar{x}_2 - 0}{s_p\sqrt{\frac{1}{n1} + \frac{1}{n2}}}$$

```
1    xbar1 = 121
2    xbar2 = 112
3
4    n1 = 10
5    n2 = 10
6
7    sigma1 = 8
8    sigma2 = 8
9
10   variance1 = sigma2^2
11   variance2 = sigma2^2
12
13   z = ((xbar1-xbar2-
14   0)/(sqrt((variance1/n1)+(variance
15   2/n2))))
16
17   alpha = 0.05
18   z.alpha = qnorm(alpha,
19   lower.tail=FALSE)
20
21
```

```
[1] "Critical Value: -2.1447866879178 2.1447866879178"
[1] "T statistics: -0.35359086434619"
```

Two Means, independent samples, unknown σ^2 (Case: unequal σ^2)

UNIVERSITI TEKNOLOGI MALAYSIA

www.utm.my

Example

Arsenic concentration in public drinking water supplies is a potential health risk. An article in the *Arizona Republic* (Sunday, May 27, 2001) reported drinking water arsenic concentrations in parts per billion (ppb) for 10 metropolitan Phoenix communities and 10 communities in rural Arizona. The data is given as follows:

INSPIRING CREATIVE AND INNOVATIVE MINDS 50

```
5
5

-xbar2-
1^2/n1)+(s2^2/n2)))

(s2^2/n2))^2/((((s1^2/

/n2)^2)/(n2-1)))

5
t(alpha/2, floor(v))
```

```
-2.16036865646279"
[1] "T statistics: -2.7744169270369"
```

Two Means, independent samples, unknown σ^2 (Case: unequal σ^2)

WHAT WE KNOW?

- Two samples:
  - Two sets of data
  - $s_1 = 7.63$, $s_2 = 15.3$
  - $n_1=10$, $n_2=10$
  - $\bar{x}_1 = 12.5$, $\bar{x}_2 = 27.5$
  - Significance: 0.05

WHAT WE WANT TO FIND?

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

$$v = \frac{\left(\frac{S_1^2}{n1} + \frac{S_2^2}{n2}\right)^2}{\frac{\left(S_1^2/n1\right)^2}{n1-1} + \frac{\left(S_2^2/n2\right)^2}{n2-1}}.$$

$$T_0^* = \frac{\bar{X}_1 - \bar{X}_2 - \Delta_0}{\sqrt{\frac{S_1^2}{n1} + \frac{S_2^2}{n2}}}$$

```
1    xbar1 = 12.5
2    xbar2 = 27.5
3
4    s1 = 7.63
5    s2 = 15.3
6
7    n1 = 10
8    n2 = 10
9
10   t0 = (xbar1-xbar2-
11   0)/(sqrt((s1^2/n1)+(s2^2/n2)))
12
13   v =
14   ((s1^2/n1)+(s2^2/n2))^2/((((s1^2/
15   n1)^2)/(n1-
16   1))+(((s2^2/n2)^2)/(n2-1)))
17
18   alpha = 0.05
19   t.alpha = qt(alpha/2, floor(v))
20
21
```

```
[1] "Critical Value: -2.16036865646279"
[1] "T statistics: -2.7744169270369"
```

THE END…

R Programming Tutorial (Part 4)

Prepared by: Chan Weng Howe

Outline

- Chi-square test using R
  - One-way contingency table
    - Categories with equal frequencies
    - Categories with unequal frequencies
  - Two-way contingency table
    - Chi-square test of independence

One-way Contingency Table (equal probabilities)

WHAT WE KNOW?

- One variable: No. of Accidents
- Categories: 5 (Mon-Fri)
- Significance level: 0.05

WHAT WE WANT TO FIND?

H_0: $P_1=P_2=P_3=P_4=P_5$
H_1: At least 1 of the 5 proportions is different from others

Function in R:
`chisq.test(observed_frequency, correct=FALSE)`

In R, chisq.test() by default applies "Yates" correction on the chi-square test. Which is arguably unnecessary, so have to set correct=FALSE

observed data

```
1   noAcc <- c(31, 42, 18, 25, 31)
2
3       output <- chisq.test(noAcc,
4
5       correct=FALSE)
6
7     Chi-squared test for given probabilities
8
9   data:  noAcc
10  X-squared = 10.653, df = 4, p-value = 0.03075
11
12
13
14
15
16
17
18
```

X^2 statistics of the observed data

P-value of the X^2 statistics

One-way Contingency Table (equal probabilities)

Function in R:

`chisq.test(observed_frequency, correction=FALSE)`

Save result into

output

observed data

```
1   noAcc <- c(31, 42, 18, 25, 31)
2
3       output <- chisq.test(noAcc,
4
5       correct=FALSE)
6
7
8
9
10
11
12
17
18
```

```
    Chi-squared test for given probabilities

data:  noAcc
X-squared = 10.653, df = 4, p-value = 0.03075
```

X^2 statistics of the ed data

P-value of the X^2 statistics

`str()` function use to inspect the structure of a variable in R

```
> str(output)
List of 9
 $ statistic: Named num 10.7
  ..- attr(*, "names")= chr "X-squared"
 $ parameter: Named num 4
  ..- attr(*, "names")= chr "df"
 $ p.value  : num 0.0308
 $ method   : chr "Chi-squared test for given probabilities"
 $ data.name: chr "noAcc"
 $ observed : num [1:5] 31 42 18 25 31
 $ expected : num [1:5] 29.4 29.4 29.4 29.4 29.4
 $ residuals: num [1:5] 0.295 2.324 -2.102 -0.811 0.295
 $ stdres   : num [1:5] 0.33 2.598 -2.351 -0.907 0.33
 - attr(*, "class")= chr "htest"
```

```
> output$statistic
X-squared
 10.65306

> output$parameter
df
 4

> output$p.value
[1] 0.03075267

> output$observed
[1] 31 42 18 25 31
> output$expected
[1] 29.4 29.4 29.4 29.4 29.4
```

Variable in R could be a "List", to access the specific information from the list can use the $ sign follow by the information name

One-way Contingency Table (equal probabilities)

$$\chi^2 = \sum \frac{(o_i - e_i)^2}{e_i}$$

WHAT WE KNOW?

- One variable: No. of Accidents
- Categories: 5 (Mon-Fri)
- Significance level: 0.05

WHAT WE WANT TO FIND?

H_0: $P_1=P_2=P_3=P_4=P_5$
H_1: At least 1 of the 5 proportions is different from others

Function in R:

qchisq(alpha, df, lower.tail=FALSE)

- To get the x2 statistics for the corresponding alpha values/p-values.

pchisq(x2, df, lower.tail=FALSE)

- To get the p-value for the corresponding x2 statistics values.

Goodness-of-Fit test in chi-square always right-tail(upper-tail), so need to set the parameter "lower.tail=FALSE"

```
1    noAcc <- c(31, 42, 18, 25, 31)
2    expprob <- sum(noAcc)/5
3    expAcc <- c(expprob, expprob,
4    expprob, expprob, expprob)
5
6    #calculate x2 for each category
7    exp <- ((noAcc-expAcc)^2)/expAcc
8
9    #sum up to a x2 statistics
10        x2 <- sum(exp)
11
12        #critical value
13        alpha <- 0.05
14        x2.alpha <- qchisq(alpha,
df=4,
16        lower.tail=FALSE)

          value of x2 of data
          ue <- pchisq(x2,

21        lower.tail=FALSE)
22
```

One-way Contingency Table (unequal probabilities)

$$\chi^2 = \sum \frac{(o_i - e_i)^2}{e_i}$$

WHAT WE KNOW?

- One variable: Frequencies of M&Ms
- Categories: 6 colors
- Significance level: 0.05
- Claimed p:

$p_1 = 0.30,\ p_2 = 0.20,\ p_3 = 0.20,\ p_4 = 0.10,\ \ p_5 = 0.10,\ p_6 = 0.10$

WHAT WE WANT TO FIND?

H_0: $p_1 = 0.30,\ p_2 = 0.20,\ p_3 = 0.20,\ p_4 = 0.10,\ \ p_5 = 0.10,\ p_6 = 0.10$

H_1: At least 1 proportions is different from claimed

Function in R:

`chisq.test(observation, p=prob, correct=FALSE)`

```
1   mnm <- c(33, 26, 21, 8, 7, 5)
2   prob <- c(0.3, 0.2, 0.2, 0.1,
3   0.1, 0.1)
4
5   # perform chi-square test on the
6   data with their correspond
7   probabilities
8   chisq.test(mnm, p=prob,
9   correct=FALSE)
10
11      #critical value
12      alpha <- 0.05
13      x2.alpha <- qchisq(alpha,
14  df=5,
15
16
17
22
```

```
> x2.alpha
[1] 11.0705
```

```
> chisq.test(mnm, p=prob, correct=FALSE)

    Chi-squared test for given probabilities

data:  mnm
X-squared = 5.95, df = 5, p-value = 0.3111
```

In R, chisq.test() by default applies "Yates" correction on the chi-square test. Which is arguably unnecessary, so have to set correct=FALSE

Since the claim has unequal probabilities, set the probabili into the parameter p of the function

Two-way Contingency Table (Test of Independence)

WHAT WE KNOW?

- Two variables:
  - Gender: Male / Female
  - Type of Crime: 4 categories
- Significance level: 0.05

WHAT WE WANT TO FIND?

H_0: Crime type is independent of gender

H_1: Crime type is not independent of gender

```
male <- c(117,150,109,124)
female <- c(66,160,168,106)
d <- data.frame(male, female)

# perform chi-square test on the
data table
chisq.test(d, correct=FALSE)

     #critical value
     alpha <- 0.05
     x2.alpha <- qchisq(alpha,
     df=3,
     lower.tail=FALSE)
```

```
> x2.alpha
[1] 7.814728
```

```
> chisq.test(d,correct=FALSE)

    Pearson's Chi-squared test

data:  d
X-squared = 28.511, df = 3, p-value = 2.837e-06
```

Two-way Contingency Table (Test of Independence)

Example 2

In the built-in data set survey, the **Smoke** column records the students smoking habit, while the **Exer** column records their exercise level.

The allowed values in Smoke are "Heavy", "Regul" (regularly), "Occas" (occasionally) and "Never". As for Exer, they are "Freq" (frequently), "Some" and "None".

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

WHAT WE WANT TO FIND?

H_0: Smoking habit is independent of the exercise level
H_1: Smoking habit is not independent of the exercise level

Function in R:
`table`(list1, list2,…)
- Group lists into a contingency table

```
1   library(MASS)
2   # get the contingency table
3   tbl = table(survey$Smoke,
4   survey$Exer)
5
6
7   # perform chi-square test on the
8   data table
9   chisq.test(tbl, correct=FALSE)
10
11      #critical val
12      alpha <- 0.05
13      x2.alpha <- qchisq(alpha,
14      df=3,
15      lower.tail=FALSE)
16
17
18
19
20
21
22
```

```
> x2.alpha
[1] 7.814728
```

```
> chisq.test(tbl,correct=FALSE)

    Pearson's Chi-squared test

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

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

Two-way Contingency Table (Test of Independence)

Example 2

In the built-in data set survey, the **Smoke** column records the students smoking habit, while the **Exer** column records their exercise level.

The allowed values in Smoke are "Heavy", "Regul" (regularly), "Occas" (occasionally) and "Never". As for Exer, they are "Freq" (frequently), "Some" and "None".

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

WHAT WE WANT TO FIND?

H_0: Smoking habit is independent of the exercise level
H_1: Smoking habit is not independent of the exercise level

```
> tbl
```

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

Warning from the chisquare could be due to the small cell values in the contingency table

```
1  library(MASS)
2  # get the contingency table
3  tbl = table(survey$Smoke,
4  survey$Exer)
5
6
7  # perform chi-square test on the
8  data table
9  chisq.test(tbl, correct=FALSE)
10
11     #critical val
12     alpha <- 0.05
13     x2.alpha <- qchisq(alpha,
14     df=3,
15     lower.tail=FALSE)
16
17
18
19
```

```
> x2.alpha
[1] 7.814728
```

```
> chisq.test(tbl,correct=FALSE)

    Pearson's Chi-squared test

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

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

Two-way Contingency Table (Test of Independence)

Example 2

In the built-in data set survey, the **Smoke** column records the students smoking habit, while the **Exer** column records their exercise level.

The allowed values in Smoke are "Heavy", "Regul" (regularly), "Occas" (occasionally) and "Never". As for Exer, they are "Freq" (frequently), "Some" and "None".

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

WHAT WE WANT TO FIND?

H_0: Smoking habit is independent of the exercise level
H_1: Smoking habit is not independent of the exercise level

```
> tbl2 = cbind(tbl[,'Freq'],tbl[,'None']+tbl[,'Some'])
> tbl2
       [,1] [,2]
Heavy     7    4
Never    87  102
Occas    12    7
Regul     9    8
```

Create another table, by merging the 2nd and 3rd columns into single columns

```
1   library(MASS)
2   # get the contingency table
3   tbl = table(survey$Smoke,
4   survey$Exer)
5   # create a new table
6   tbl2 = cbind(tbl[,'Freq'],
7   tbl[,'None']+tbl[,'Some'])
8
9   # perform chi-square test on the
10  data table
11  chisq.test(tbl2, correct=FALSE)
12
13      #critical val
14      alpha <- 0.05
15      x2.alpha <- qchisq(alpha,
16      df=3,
17      lower.tail=FALSE)
18
19
20
21
22
```

```
> x2.alpha
[1] 7.814728
```

```
> chisq.test(tbl2,correct=FALSE)

    Pearson's Chi-squared test

data:  tbl2
X-squared = 3.2328, df = 3, p-value = 0.3571
```

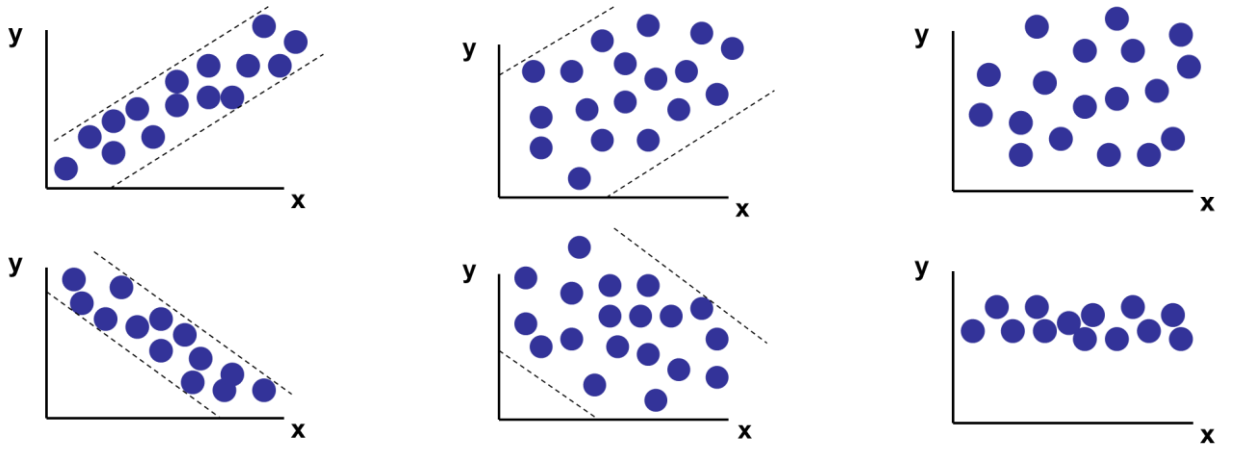
THE END…

R Programming Tutorial

Prepared by: Chan Weng Howe

Correlation Analysis

- Measure the **strength** of relationship between **two** variables.

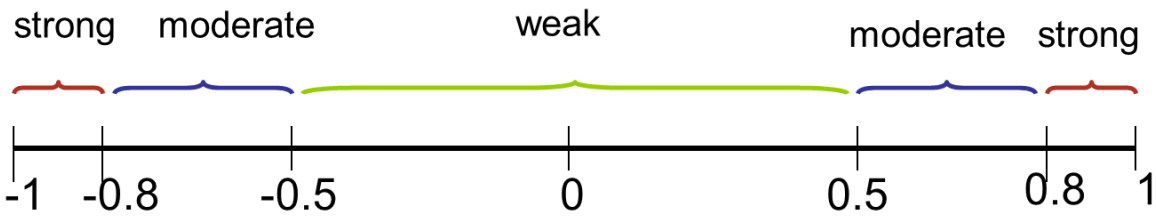


Strong relationship

Weak relationship

No relationship

Strength of correlation coefficient (r)



Correlation analysis in R

- Correlation efficient (r) values can be calculated in R via **cor()** function.
- How to use? **cor(x,y)**
  - The default method used is **Pearson's**
- Example data:

Students	Study Hours	GPA
S1	42	3.3
S2	23	2.9
S3	31	3.2
S4	35	3.2
S5	16	1.9
S6	26	2.4
S7	39	3.7
S8	19	2.5

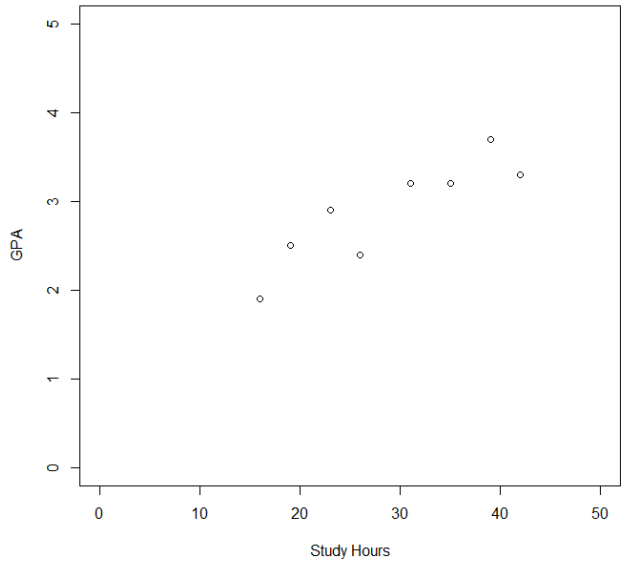
```
x <- c(42,23,31,35,16,26,39,19)
y <- c(3.3, 2.9, 3.2, 3.2, 1.9,
2.4, 3.7, 2.5)

    #calculate corr. coefficient
    cor(x,y)
```

```
> cor(x,y)
[1] 0.8837091
```

Correlation analysis in R

- How we can view the data in graph?
  via **Scatter plot**
- Use simple **plot()** function in R.




```
x <- c(42,23,31,35,16,26,39,19)
y <- c(3.3, 2.9, 3.2, 3.2, 1.9,
2.4, 3.7, 2.5)

      #calculate corr. coefficient
      cor(x,y)

      plot(x,y, xlim=c(0,50),
      ylim=c(0,5), xlab="Study
Hours",  ylab="GPA")
```

Correlation analysis in R

- More example:

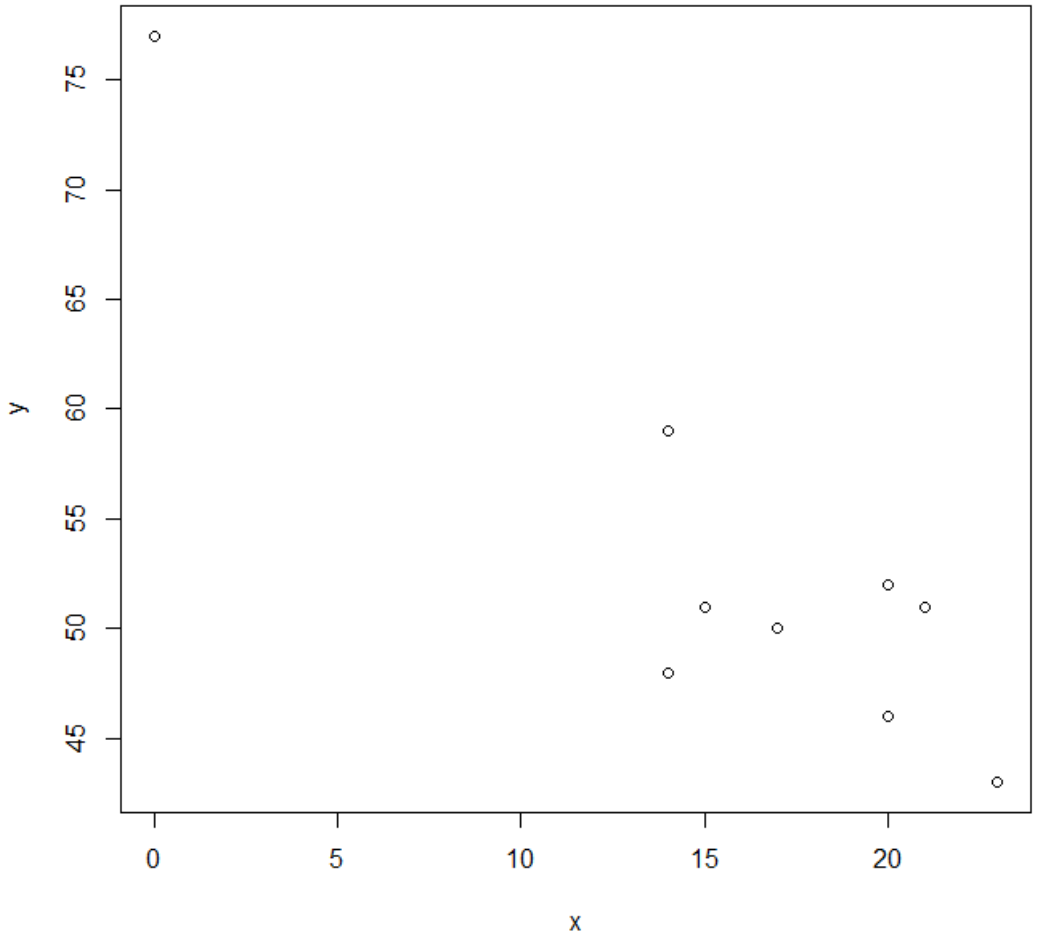
x	23	14	14	0	17	20	20	15	21
y	43	59	48	77	50	52	46	51	51

The data represent x=score on a measure of test anxiety and y=exam score for a sample of n=9 students:

a) Construct a scatter plot, and comment on the features of the plot.
b) Does there appear to be a linear relationship between the two variables? How would you characterize the relationship?
c) Compute the value of the correlation coefficient. Is the value of r consistent with your answer to part (b)?
d) Is it reasonable to conclude that test anxiety caused poor exam performance? Explain.

```
1    x<- c(23,14,14,0,17,20,20,15,21)
2    y<- c(43,59,48,77,50,52,46,51,51)
3
4    plot(x,y)
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
```

Correlation analysis in R



Correlation analysis in R

- More example:

x	23	14	14	0	17	20	20	15	21
y	43	59	48	77	50	52	46	51	51

The data represent x=score on a measure of test anxiety and y=exam score for a sample of n=9 students:

a) Construct a scatter plot, and comment on the features of the plot.
b) Does there appear to be a linear relationship between the two variables? How would you characterize the relationship?
c) Compute the value of the correlation coefficient. Is the value of r consistent with your answer to part (b)?
d) Is it reasonable to conclude that test anxiety caused poor exam performance? Explain.

```
1   x<- c(23,14,14,0,17,20,20,15,21)
2   y<- c(43,59,48,77,50,52,46,51,51)
3
4   plot(x,y)
5
6   cor(x,y)
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
```

```
> cor(x,y)
[1] -0.9124331
```

Correlation analysis in R

- To perform linear regression in R, use the **lm()** function
- This will build a **linear regression model** based on the data
- How to use the function:
  - Assume you have your x, y
  - Use: **lm(y~x)**
  - Expression of "~" in R means it is a formula. Where y~x is equivalent to y = f(x)

```
1   x <- c(1400, 1600, 1700, 1875,
2   1100, 1550, 2350, 2450, 1425,
3   1700)
4
5   y <- c(245, 312, 279, 308, 199,
6   219, 405, 324, 319, 255)
7
8   model <- lm(y~x)
9   model
10
11
12
13
14
15
16
17
18
19
20
21
22
```

```
> model

Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x
    98.2483       0.1098
```

Linear regression model

$$\hat{y} = 98.2483 + 0.1098x$$

Correlation analysis in R

- To perform linear regression in R, use the **lm()** function
- This will build a **linear regression model** based on the data
- How to use the function:
  - Assume you have your x, y
  - Use: **lm(y~x)**
  - Expression of "~" in R means it is a formula. Where y~x is equivalent to y = f(x)

```
1   x <- c(1400, 1600, 1700, 1875,
2   1100, 1550, 2350, 2450, 1425,
3   1700)
4
5   y <- c(245, 312, 279, 308, 199,
6   219, 405, 324, 319, 255)
7
8   model <- lm(y~x)
9   model
10
11  summary(model)
12
13
14
15
16
17
18
19
20
21
```

Linear regression model

$$\hat{y} = 98.2483 + 0.1098x$$

Regression in R

Try adjust the plot with custom limits for x-axis and y-axis

```
> summary(model)

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-49.388 -27.388  -6.388  29.577  64.333

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 98.24833   58.03348   1.693   0.1289
x            0.10977    0.03297   3.329   0.0104 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 41.33 on 8 degrees of freedom
Multiple R-squared:  0.5808,    Adjusted R-squared:  0.5284
F-statistic: 11.08 on 1 and 8 DF,  p-value: 0.01039
```

$$\hat{y} = 98.2483 + 0.1098x$$

Correlation analysis in R

- What if we wan to see the plot of the model we learn from the data?
- First, plot a scatter plot based on x and y using **plot()** function
- Add the linear regression model into the plot using **abline()** function

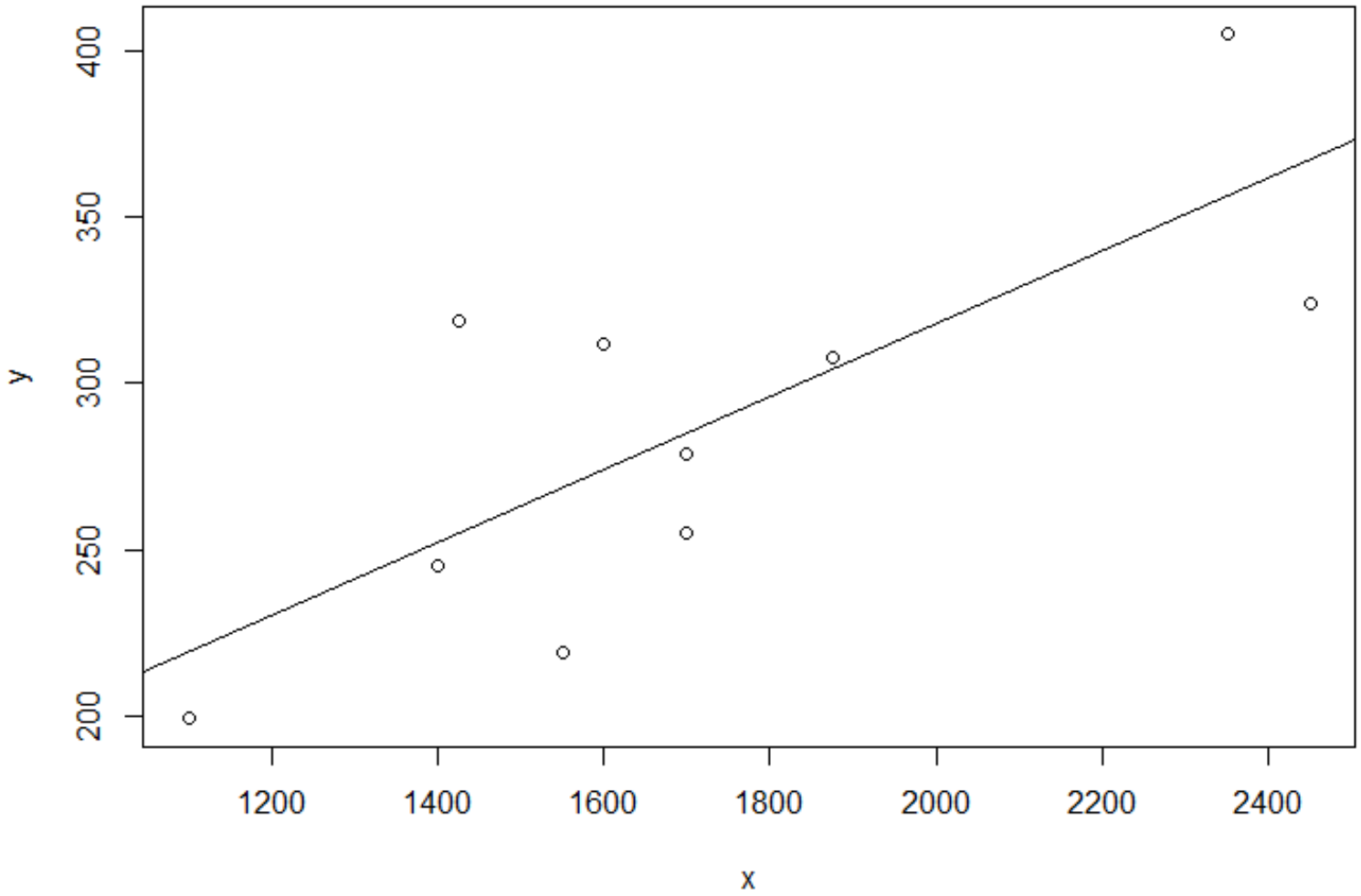
```
1   x <- c(1400, 1600, 1700, 1875,
2   1100, 1550, 2350, 2450, 1425,
3   1700)
4
5   y <- c(245, 312, 279, 308, 199,
6   219, 405, 324, 319, 255)
7
8   model <- lm(y~x)
9   model
10
11  plot(x,y)
12  abline(model)
13
14
15
16
17
18
19
20
21
22
```

Linear regression model

$$\hat{y} = 98.2483 + 0.1098x$$

Regression in R

Try adjust the plot with custom limits for x-axis and y-axis



Regression in R

Replotting using:

```
plot(x,y, xlim=c(0,2600), ylim=c(0,500))
abline(model)
```