

Essentials of Data Science With R Software - 1

Probability and Statistical Inference

Lecture 1

Data Science- Why, What and How?

Shalabh

Department of Mathematics and Statistics

Indian Institute of Technology Kanpur

What is Data Science?

What is Science?

- Special approach to find the answer of a query.
- We want to know the reason - How, Why, Where, Who...?

What is Data?

- Source of reliable information.

What is Data Science?

- A scientific approach of retrieving the information from data.

Why data is needed?

Why do we collect the data?

To quench the thirst for knowledge.

We want to know the reason - How, Why, Where, Who...?

Easiest option- get some data about the relevant question.

Answer all the questions on the basis of collected data.

Data and Statistics

Data: Very important source of information but cannot speak itself.

We cannot understand what data is telling us.

Statistics is the language of data.

How to collect the data, how to analyze that, how to draw correct statistical inferences, how to decide for the correct statistical tool on the numerical facts is referred as data analysis.

Statistics is a science of turning data into information to be used for decision making.

Data and Statistics

Proper interpretation of inferences is important.

Statistics can not do miracles.

Statistics can not change the process or phenomenon.

Its a scientific way of extracting and retriving information.

Why collect data?

- **To verify theoretical findings,**
- **Draw inferences just on the basis of collected data,**
- **Developing statistical models, which can be further used for policy decisions, classification, forecasting etc.**

Data and Statistics

Statistics is a language of data.

	Correct data	Wrong data
Correct statistical tool	Correct decision	Incorrect decision
Wrong statistical tool	Incorrect decision	Incorrect decision

Rule: Garbage in – Garbage out

Statistics has its own derived rules.

Rules are framed such that correct decisions, as indicated by the data and based on the hidden information, are taken.

It does forecasting but not like astrologer's parrot.

Statistics and Data Science

How Statistics got transformed to Data Science?

What is expected from Data Science which was not expected from “Statistics”.

Advent and rapid development in computers have impacted Statistics.

Earlier, it was difficult to collect the data and even many times the data was not available.

Now, data is easily available and too much data is available.

Big data analysis is the latest news, petabyte is the unit of data size.

Statistics, Computers and Data Science

Earlier, the emphasis was on theoretical developments in Statistics.

Computers helped in the development of from “Computational Statistics”.

If theory and mathematical analysis became complicated, the computational statistics supplemented it.

With the computational support , the theoretical developments in statistics gained more relevance and applications.

The computations and statistics became the two inseperable parts of data science.

Statistics, Computers and Data Science

Once we adventure into the Computational Statistics, the role and use of computers became very important.

Computers require programming language, software, data management and several other aspects.

The areas of applications of statistics have increased.

Topics like artificial intelligence, machine learning, supervised learning, unsupervised learning, reinforcement learning are based on statistics but they are heavily based on computers.

Statistics, Computer Science and Data Science

Data science has various ingredients- Statistics, mathematics, computer science, ...

Objectives of statistics and data science are the same.

Statistics aims to extract the information contained in the data and so is the aim of data science.

Data science, when applied to different fields can lead to incredible new insights.

Statistics, Computer Science and Data Science

The only form of data that matters in decision science is digital data.

Digital data is information that is not so easily interpretable by an individual. It depends upon machines to interpret/ process and/or alter it.

What we see on a computer screen – text, photo, movie etc., they are the digital letters which is essentially a systematic collection of coded ones and zeros.

Expectation from Data Scientist

What is needed to become the data scientist?

First decide what we want to become- A Doctor or a Compounder?

Decide-

Want to only use the tools?

Want to understand the utility of tools?

Or want to develop the tools?

In my opinion- all are needed.

Role of Statistics in Data Science

Statistics is the soul of data science.

• Descriptive statistics	• Nonparametric inference
• Probability theory	• Multivariate analysis
• Statistical inference	• Linear regression analysis
• Decision theory	• Nonlinear regression analysis
• Bayesian inference	• Simulation techniques
• Frequentist inference	• Monte Carlo methods
• Parametric inference	•

Role of Statistics in Data Science

The theoretical developments are essential which are needed to be exposed to computational procedures.

Computational procedures have their own limitations and so optimization methods are required.

The implementation of statistical, mathematical, optimization methods etc. are to be simultaneously implemented over a data set and for that, data management is required.

All these aspects are logically implemented in a systematic way and correct statistical inferences are drawn.

Role of Statistics in Data Science

Based on the obtained inferences, proper interpretations are made and used for policy formulation, policy prescription and further applications like forecasting etc.

Without learning the basic tools of Statistics, it is not possible to learn data science. So proper knowledge of all the fields is required to become a data scientist.

My role?

My job?

Essentials of Data Science With R Software - 1

Probability and Statistical Inference

Introduction to R Software

:::

Lecture 2

Installation and Working with R

Shalabh

Department of Mathematics and Statistics

Indian Institute of Technology Kanpur

R Software

Use of a software is desirable and moreover an essential part of any analysis.

Some popular statistical software are SPSS, SAS, Minitab, Stata, Matlab etc.

Another software is R.

R is a free software.

Developers of R Software

Currently developed by the R Development Core Team.

Available at www.r-project.org

It supports many free packages which helps the data scientist and analyst.

What is R?

R is an environment for data manipulation, statistical computing, graphics display and data analysis.

Effective data handling and storage of outputs is possible.

Simple as well as complicated calculations are possible.

Simulations are possible.

What is R?

Graphical display on-screen and hardcopy are possible.

Programming language is effective which includes all possibilities just like any other good programming language.

R has a statistical computing environment.

R is free (open source) software and therefore is not a black box.

Switching to R

R is available for Windows, Unix, Linux and Macintosh platforms.

Built in and contributed packages are available, and users are provided tools to make packages.

It is possible to contribute own packages.

The commands can be saved, run and stored in script files.

Graphics can be directly saved in a Postscript or PDF format.

Installing R

You may install R in a windows or Mac platform by downloading from the Comprehensive R Archive Network (CRAN) website: www.r-project.org or directly from <http://cran.r-project.org/>

https://www.r-project.org

133%

...

🔒

☆

🔍



[Home]

Download

CRAN

R Project

About R

Logo

Contributors

What's New?

Reporting Bugs

The R Project for Statistical Computing

Getting Started

R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS.

To **download R**, please choose your preferred [CRAN mirror](#).

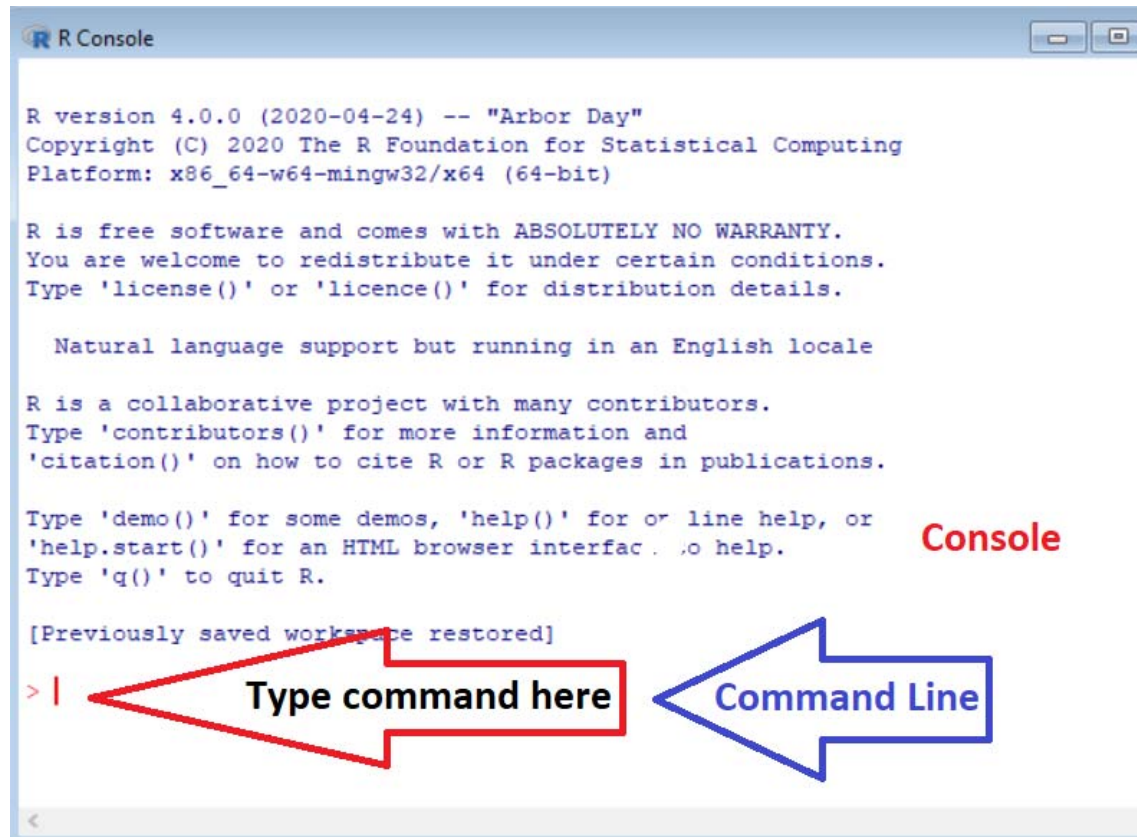
If you have questions about R like how to download and install the software, or what the license terms are, please read our [answers to frequently asked questions](#) before you send an email.

Installing R



Icon will appear.

Double click on this icon will start the software.

A screenshot of the R Console window. The title bar says "R Console". The text inside the window is as follows:

```
R version 4.0.0 (2020-04-24) -- "Arbor Day"
Copyright (C) 2020 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[Previously saved workspace restored]
> |
```

Two arrows point to the prompt line. A red arrow points from the left with the text "Type command here". A blue arrow points from the right with the text "Command Line". The word "Console" is written in red text to the right of the prompt line.

Installing Packages and Libraries

The base R package contains programs for basic operations.

It does not contain some of the libraries necessary for advanced statistical work.

Specific requirements are met by special packages.

They are downloaded and their downloading is very simple.

Installing Packages and Libraries

The base R package contains some necessary libraries only.

Other libraries are required for advanced statistical work which are downloaded and installed as and when required.

Run the R program, then use the `install.packages` command to download the libraries.

Examples :

`install.packages("ggplot2")` : installs package `ggplot2`

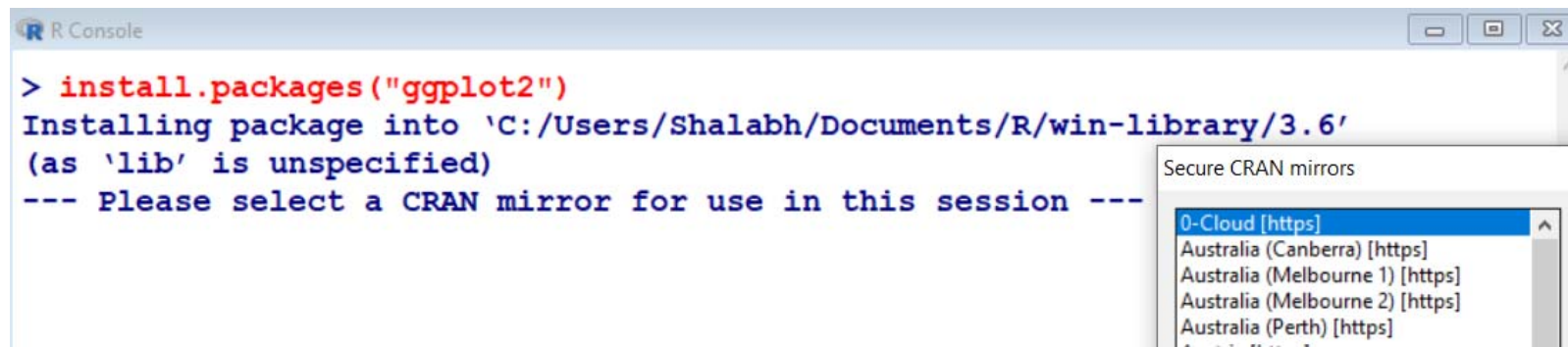
`install.packages("agricolae")` : installs package `agricolae`

`install.packages("DoE.base")` : installs package `DoE.base`

Installing Packages and Libraries

Example

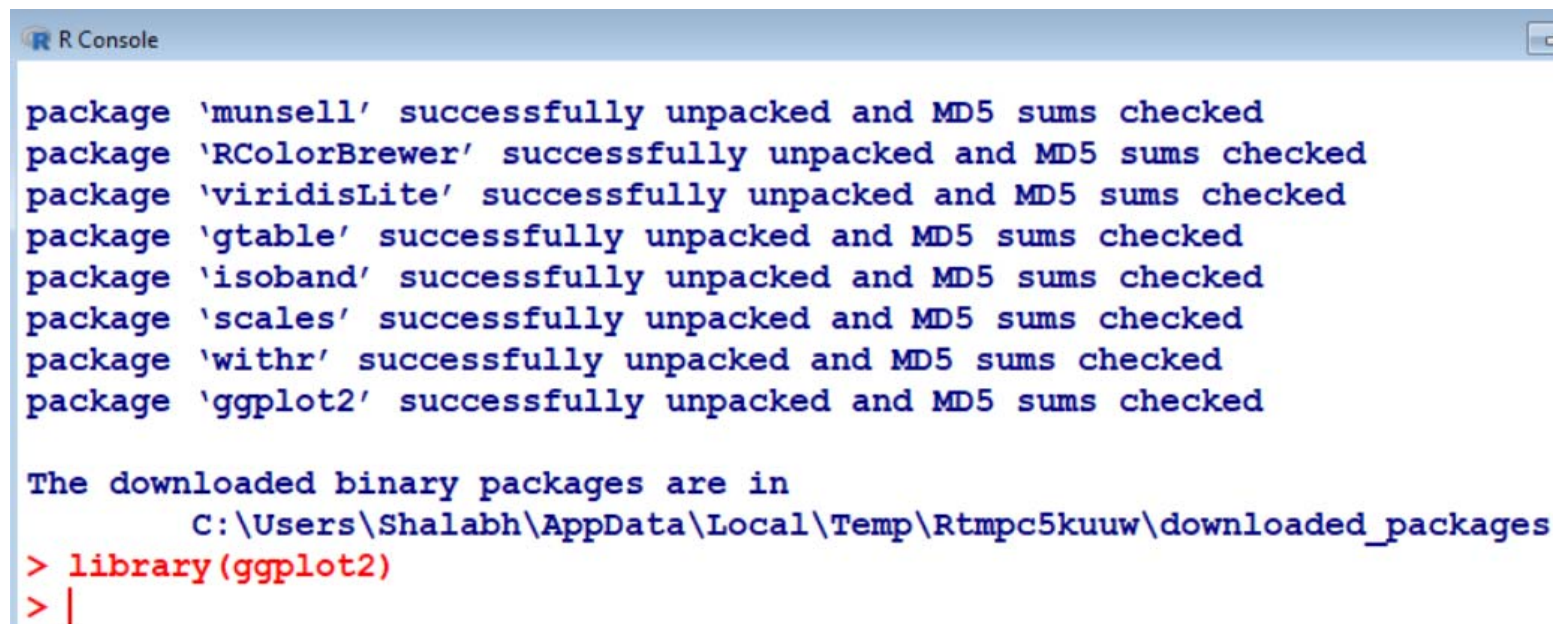
```
install.packages("ggplot2")
```



The image shows an R Console window with the following text:

```
> install.packages("ggplot2")  
Installing package into 'C:/Users/Shalabh/Documents/R/win-library/3.6'  
(as 'lib' is unspecified)  
--- Please select a CRAN mirror for use in this session ---
```

A dialog box titled "Secure CRAN mirrors" is open, showing a list of mirrors. The first mirror, "0-Cloud [https]", is selected and highlighted in blue. Other visible mirrors include "Australia (Canberra) [https]", "Australia (Melbourne 1) [https]", "Australia (Melbourne 2) [https]", and "Australia (Perth) [https]".



The image shows an R Console window with the following text:

```
package 'munsell' successfully unpacked and MD5 sums checked  
package 'RColorBrewer' successfully unpacked and MD5 sums checked  
package 'viridisLite' successfully unpacked and MD5 sums checked  
package 'gtable' successfully unpacked and MD5 sums checked  
package 'isoband' successfully unpacked and MD5 sums checked  
package 'scales' successfully unpacked and MD5 sums checked  
package 'withr' successfully unpacked and MD5 sums checked  
package 'ggplot2' successfully unpacked and MD5 sums checked  
  
The downloaded binary packages are in  
  C:\Users\Shalabh\AppData\Local\Temp\Rtmpc5kuuw\downloaded_packages  
> library(ggplot2)  
> |
```

Libraries in R

To use a library, type the `library` function with the name of the library in brackets.

Thus to load the `ggplot2` library type:

```
library(ggplot2)
```

Similarly,

```
library(agricolae) : loads package agricolae
```

```
library(DoE.base) : loads package DoE.base
```

Libraries in R

Examples of libraries that come as a part of base package in R.

MASS : package associated with Venables and Ripley's book entitled *Modern Applied Statistics using S-Plus*.

library(MASS) loads package **MASS**

Contents of Libraries

Use `help` function to get the detailed contents of library packages.

We find out about the contents of the `agricolae` library using

`library(help=agricolae)` command

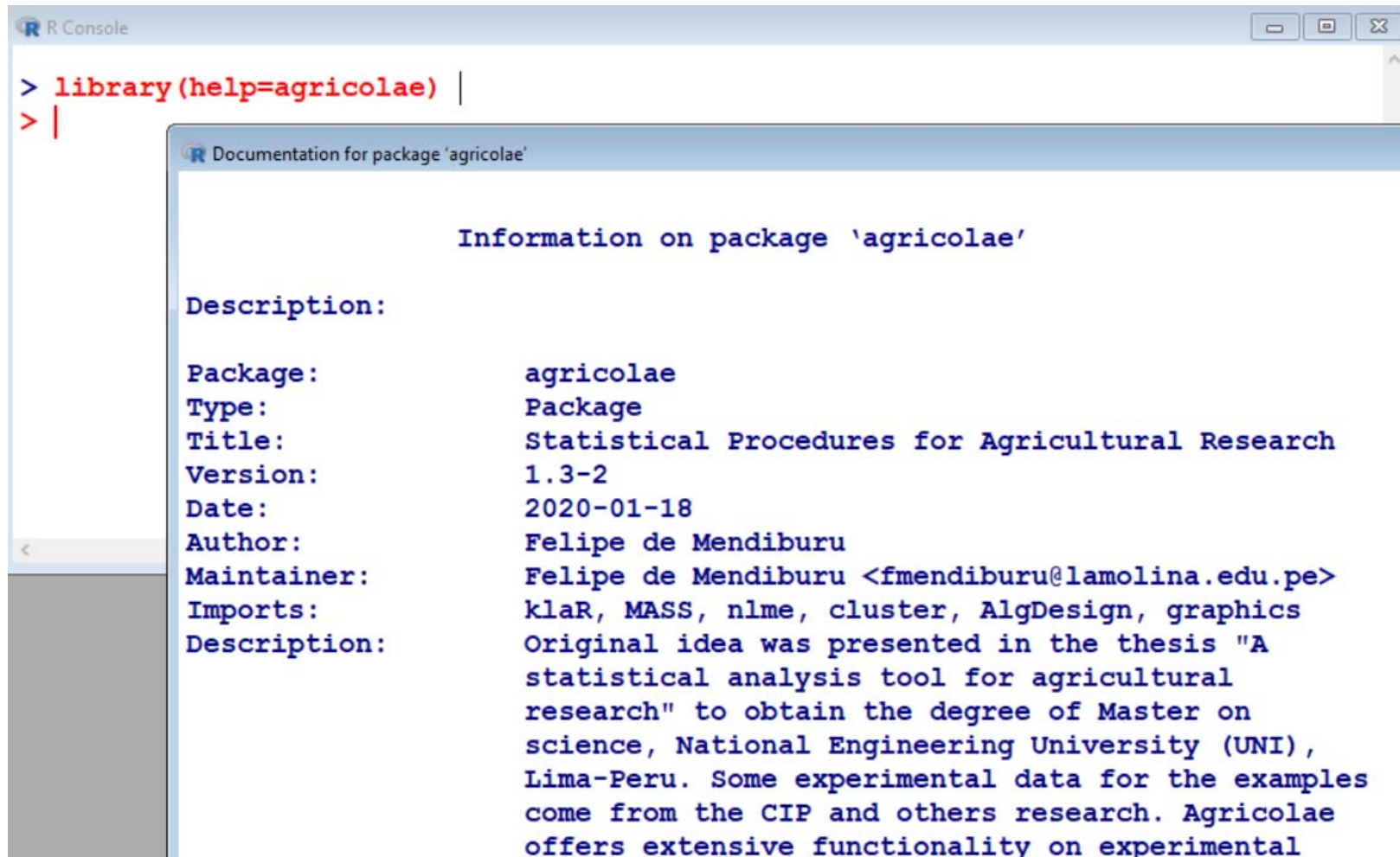
```
Information on package 'agricolae'
```

```
Description:
```

```
Package:      agricolae
Type:         Package
Title:        Statistical Procedures for Agricultural Research
Version:      1.3-2
Date:         2020-01-18
Author:       Felipe de Mendiburu
Maintainer:   Felipe de Mendiburu <fmendiburu@lamolina.edu.pe>
Imports:      klaR, MASS, nlme, cluster, AlgDesign, graphics
Description:  Original idea was presented in the thesis "A
              statistical analysis tool for agricultural ... ..
... ..
```

followed by a list of all the functions and data sets.

Contents of Libraries



```
R Console
> library(help=agricolae)
> |
```

Documentation for package 'agricolae'

Information on package 'agricolae'

Description:

Package:	agricolae
Type:	Package
Title:	Statistical Procedures for Agricultural Research
Version:	1.3-2
Date:	2020-01-18
Author:	Felipe de Mendiburu
Maintainer:	Felipe de Mendiburu <fmendiburu@lamolina.edu.pe>
Imports:	klaR, MASS, nlme, cluster, AlgDesign, graphics
Description:	Original idea was presented in the thesis "A statistical analysis tool for agricultural research" to obtain the degree of Master on science, National Engineering University (UNI), Lima-Peru. Some experimental data for the examples come from the CIP and others research. Agricolae offers extensive functionality on experimental

Cleaning up the Windows

We assign names to variables when analyzing any data.

It is good practice to remove the variable names given to any data frame at the end each session in R.

`rm()` command removes variable names

For example,

`rm(x,y,z)` removes the variables `x`, `y` and `z`.

How to clear the screen in R

Press **ctrl + L** to clear the screen of R console.

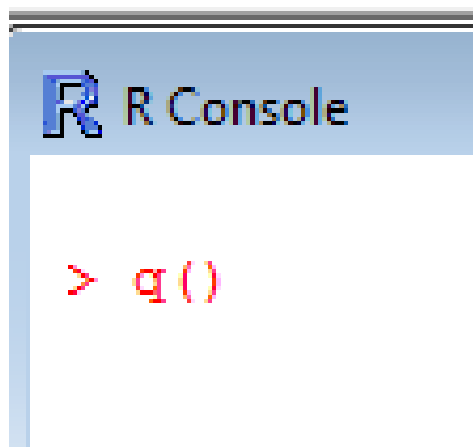


and



How to quit in R

Type **q()** to quit R.



Essentials of Data Science With R Software - 1

Probability and Statistical Inference

Introduction to R Software

:::

Lecture 3

Installation and Working with R Studio

Shalabh

Department of Mathematics and Statistics

Indian Institute of Technology Kanpur

Working with R Studio

Use command line to type and execute the commands.

Some free software like R Studio, Tinn R etc. are also available to work with R software.

They are the interface between R and us and help in running the R software. Such software make coding and execution of programmes easier.

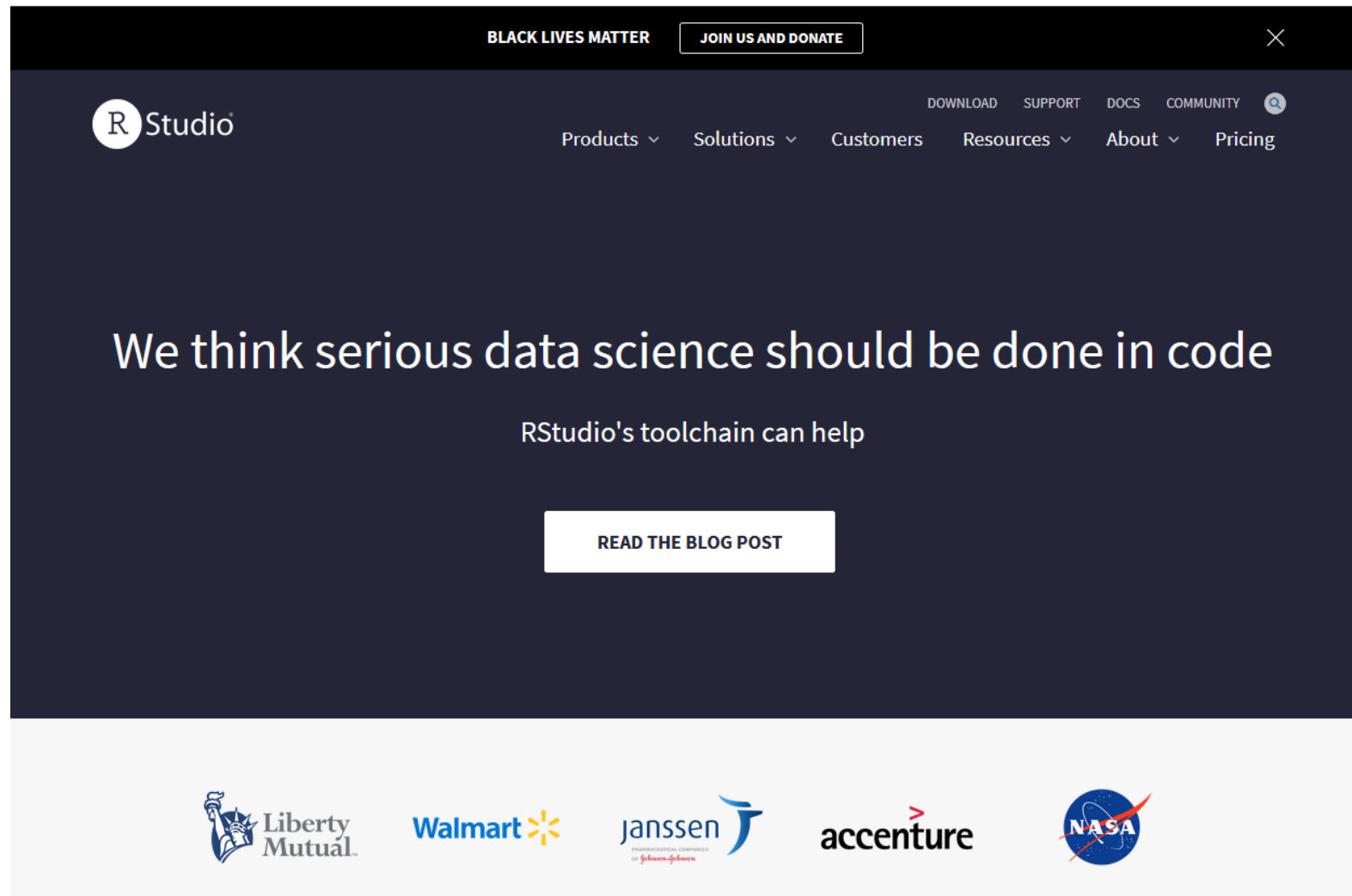
R Studio is available at <https://www.rstudio.com/>

Rstudio is written in C++ programming language.

Rstudio is a free and open-source integrated development environment (IDE) for R.

Tinn R is available at <https://sourceforge.net/projects/tinn-r/>

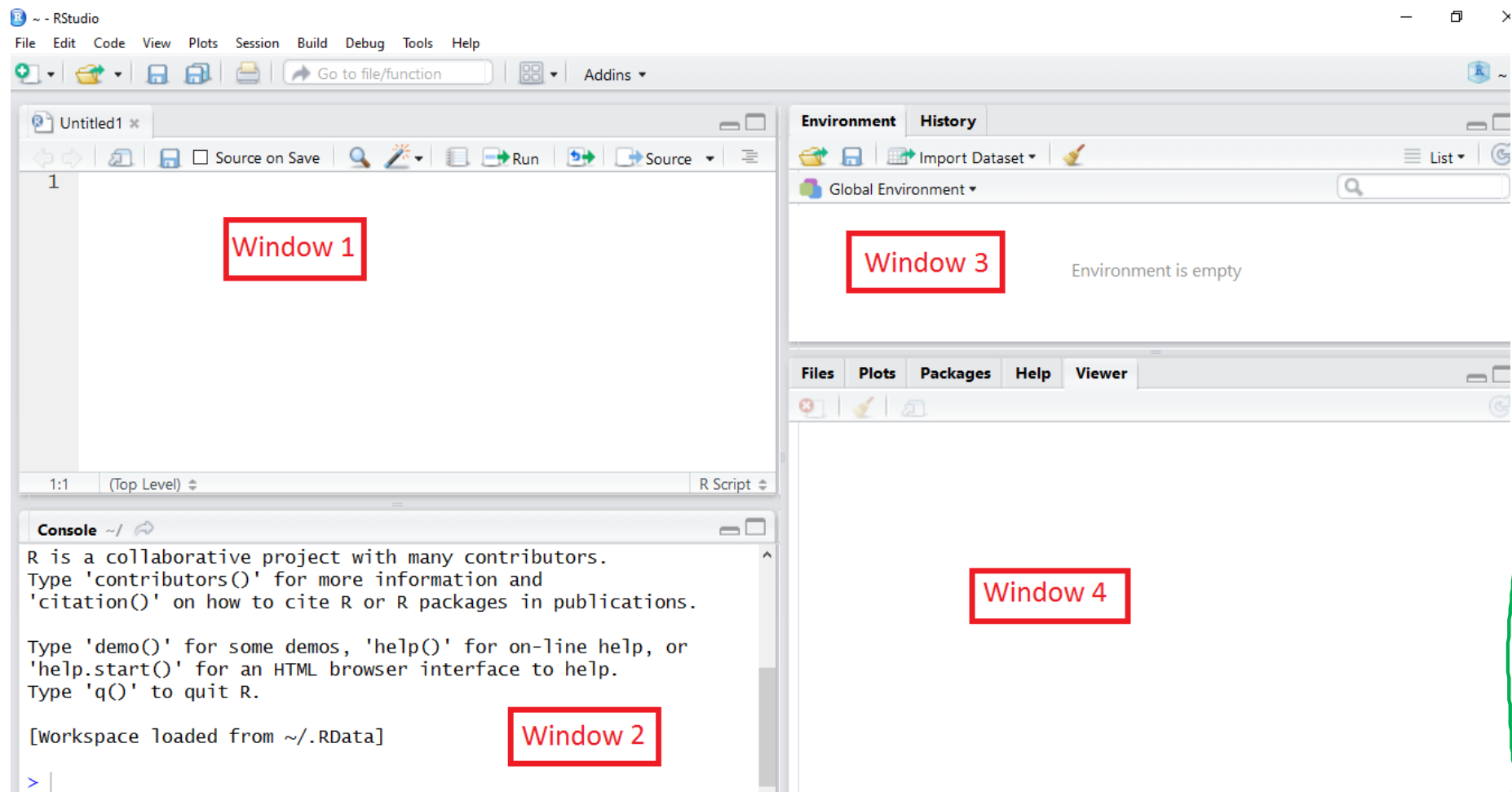
Installing R Studio



Download and double click on the downloaded file.

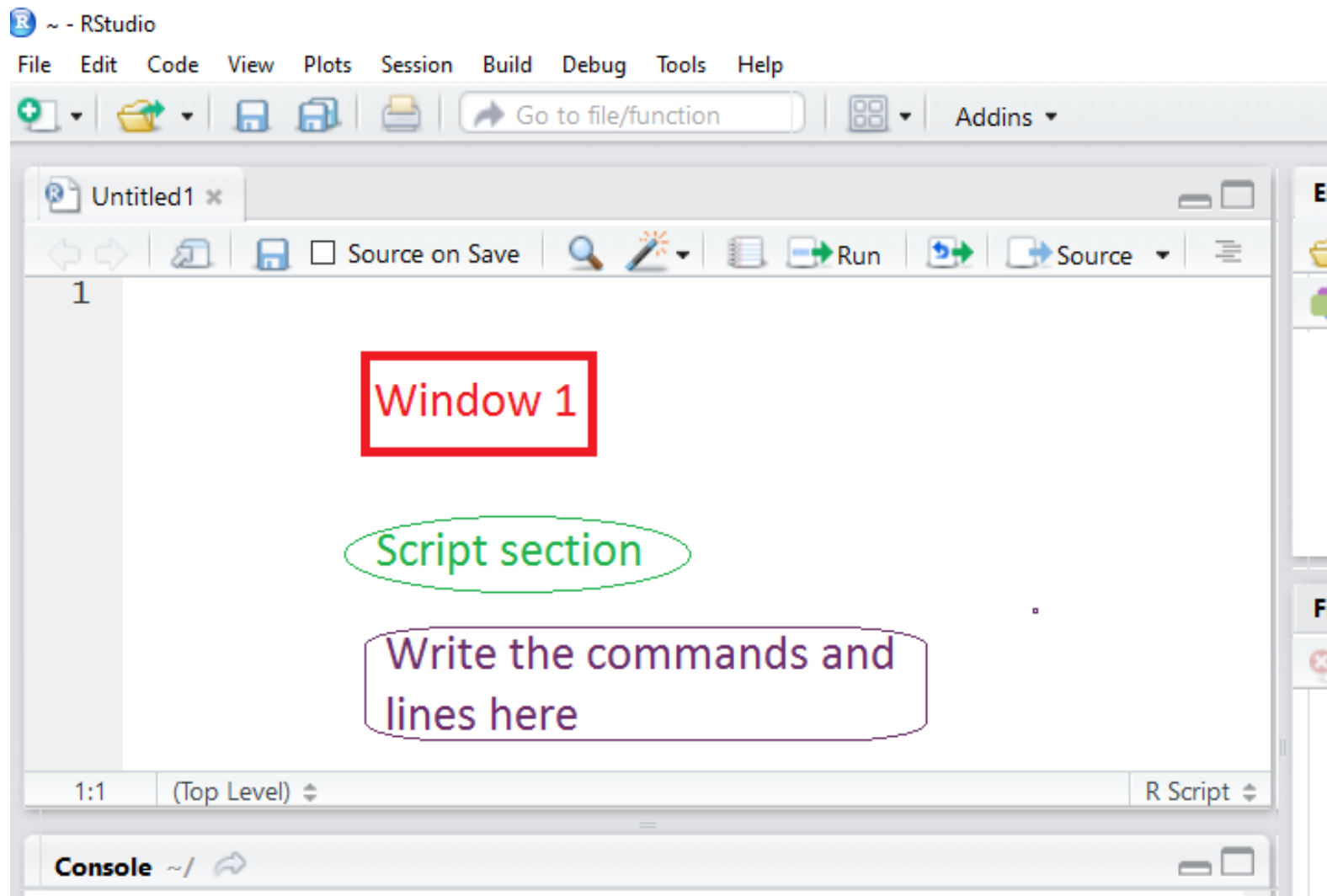
Introduction to R Studio

First opening window of R Studio is as follows having four windows.



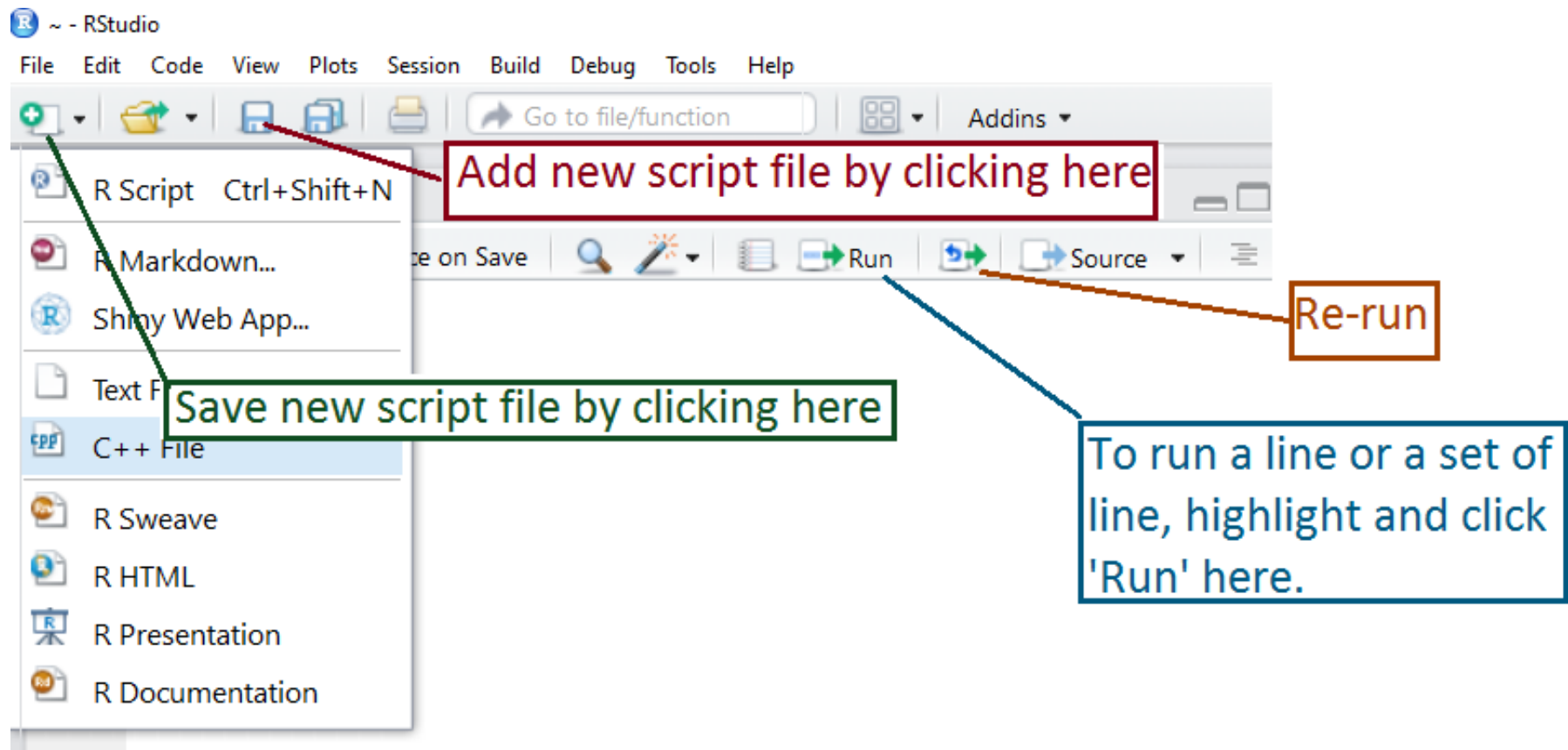
Introduction to R Studio

Description of Window 1



Introduction to R Studio

Description of Window 1



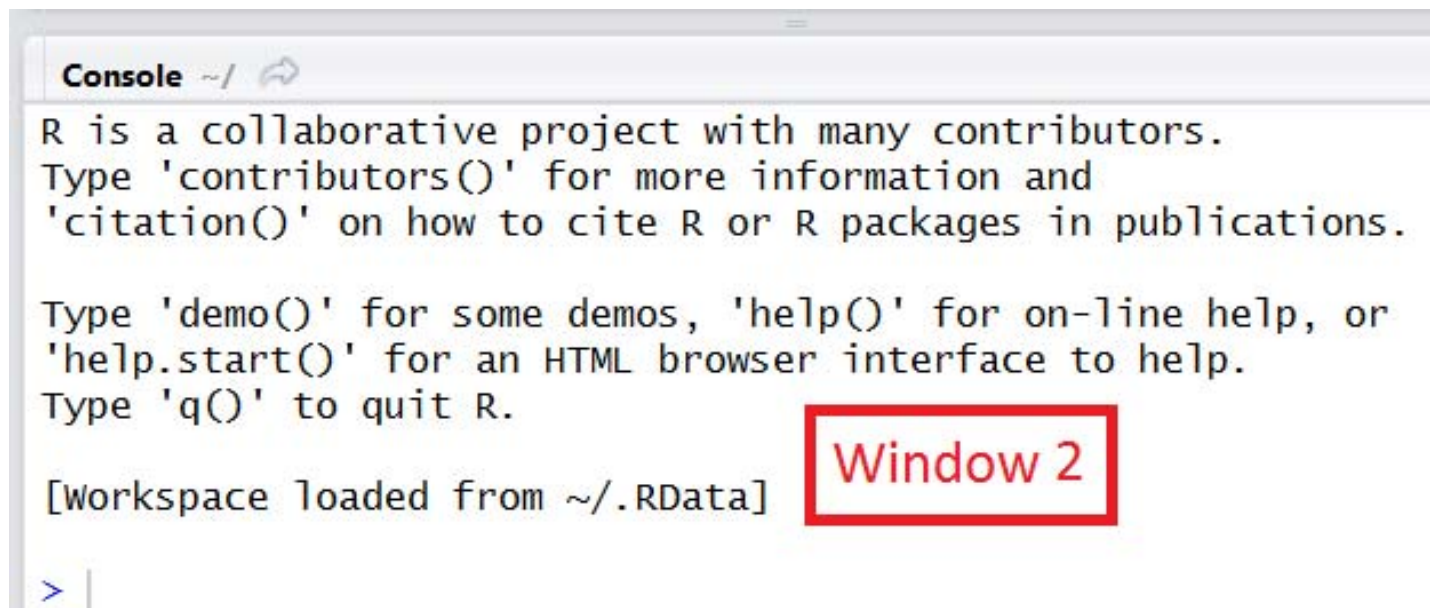
Introduction to R Studio


Description of Window 2 : Console

R program window appears here.

Calculations take place in console window.

One can write programmes in console also but it is hard to make corrections and experiments with the coding.



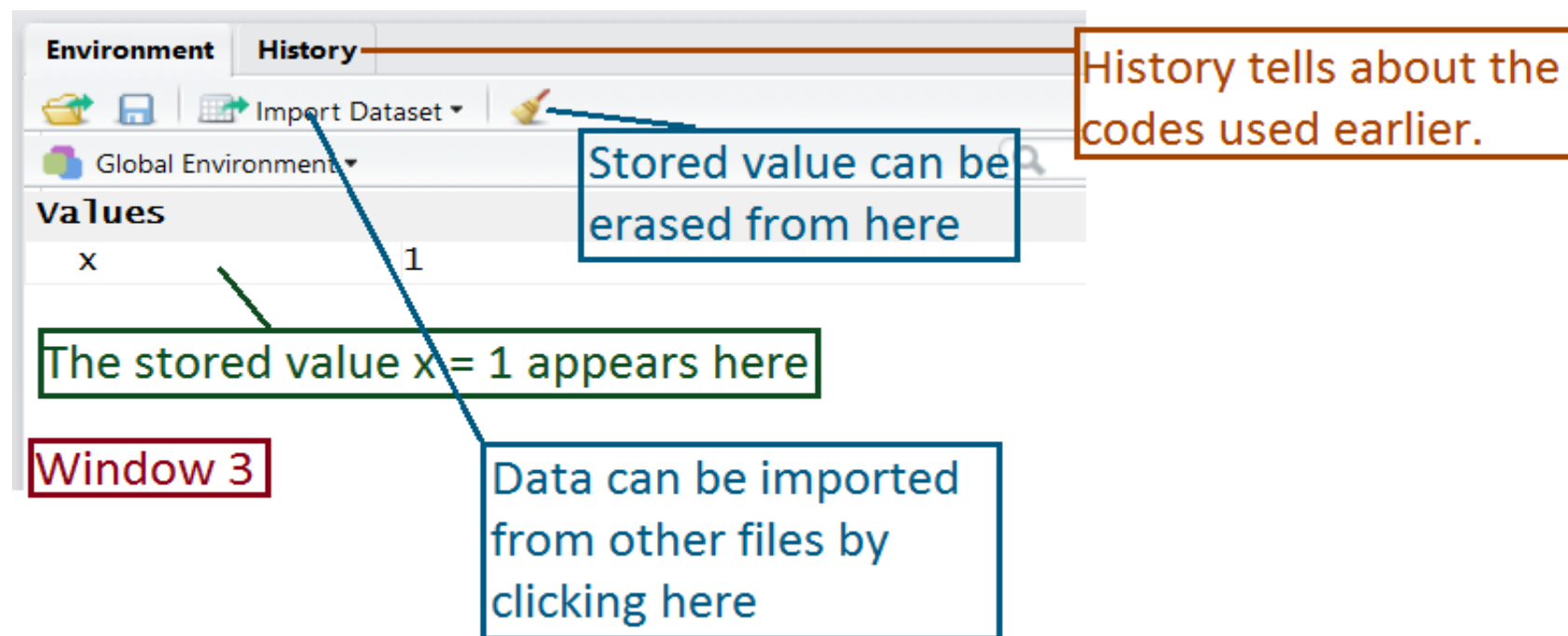
```
Console ~/   
R is a collaborative project with many contributors.  
Type 'contributors()' for more information and  
'citation()' on how to cite R or R packages in publications.  
  
Type 'demo()' for some demos, 'help()' for on-line help, or  
'help.start()' for an HTML browser interface to help.  
Type 'q()' to quit R.  
  
[Workspace loaded from ~/.RData]  
> |
```

Window 2

Introduction to R Studio

Description of Window 3 : Environment window

All the variables and objects used in the programme appear here.
The nature and values of variables and objects also appear here.



The screenshot shows the R Studio Environment window. At the top, there are two tabs: 'Environment' and 'History'. Below the tabs is a toolbar with icons for file operations and a button labeled 'Import Dataset'. The main area is divided into two sections: 'Global Environment' and 'Values'. The 'Values' section displays a table with two columns: 'x' and '1'. A green box with a green border contains the text 'The stored value x = 1 appears here', with a green arrow pointing to the 'x' column. A blue box with a blue border contains the text 'Stored value can be erased from here', with a blue arrow pointing to the '1' column. A red box with a red border contains the text 'History tells about the codes used earlier.', with a red arrow pointing to the 'History' tab. A blue box with a blue border contains the text 'Data can be imported from other files by clicking here', with a blue arrow pointing to the 'Import Dataset' button. A red box with a red border contains the text 'Window 3'.

Environment History

Import Dataset

Global Environment

Values

x 1

History tells about the codes used earlier.

Stored value can be erased from here

The stored value x = 1 appears here

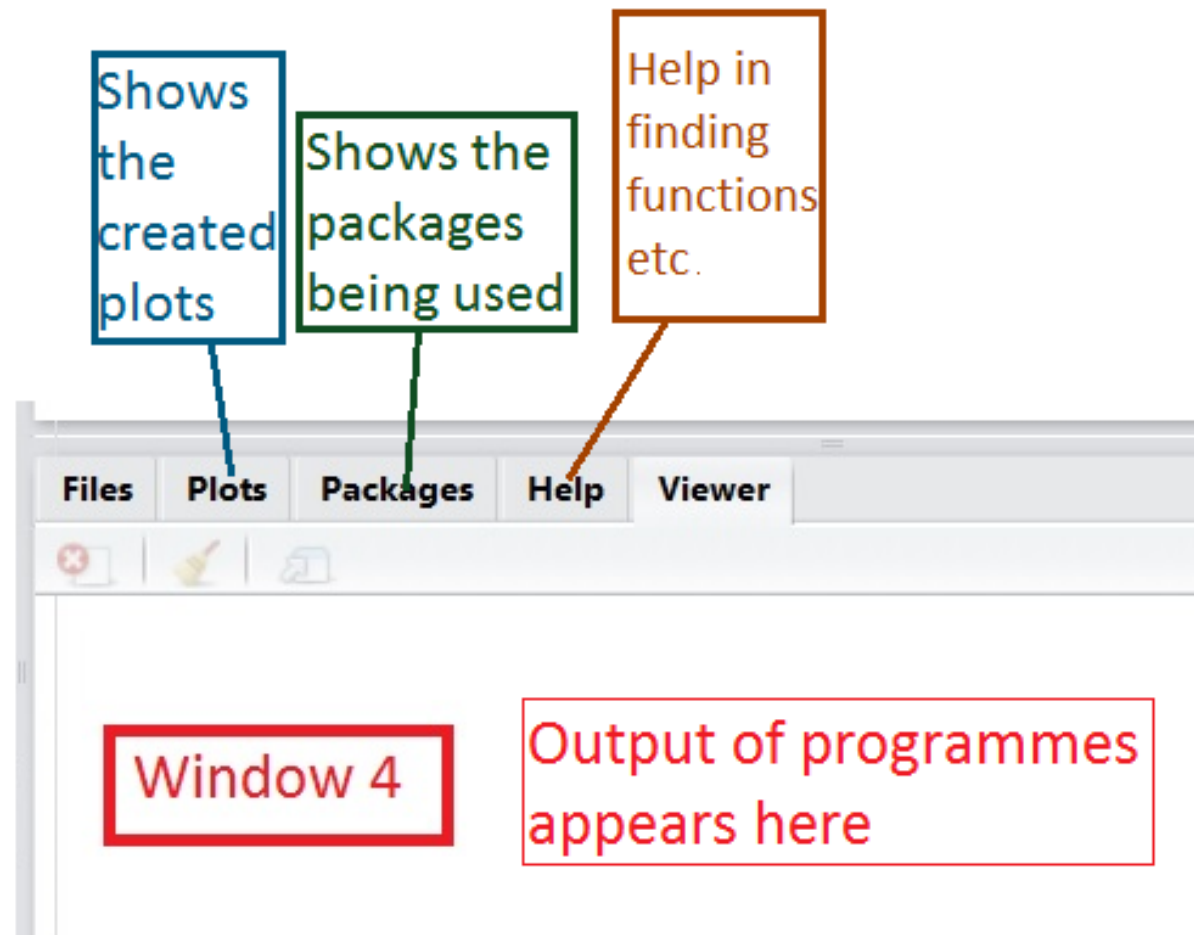
Window 3

Data can be imported from other files by clicking here

Introduction to R Studio

Description of Window 4 : Output window

The output of programmes appears in this window.



Introduction to R Studio

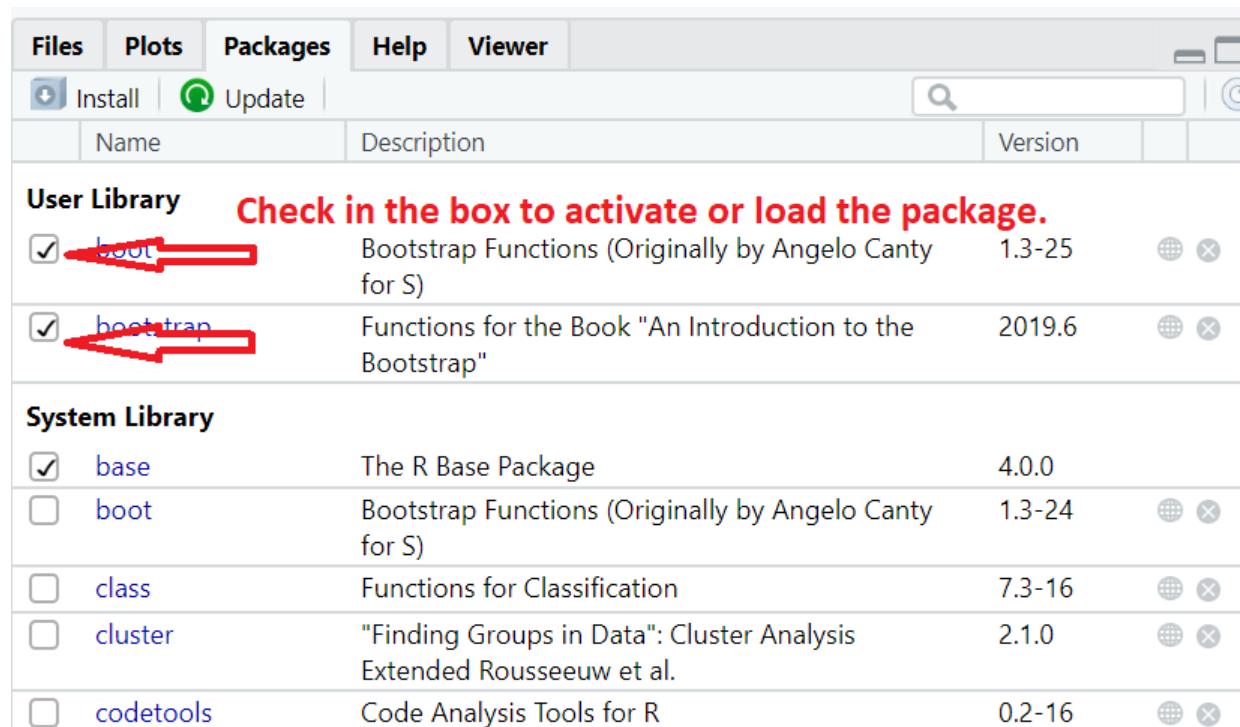
Description of Window 4 : Output window

Packages:

All the packages being installed appear here.

Packages are not active.

Check mark in the boxes to activate them.



Introduction to R Studio

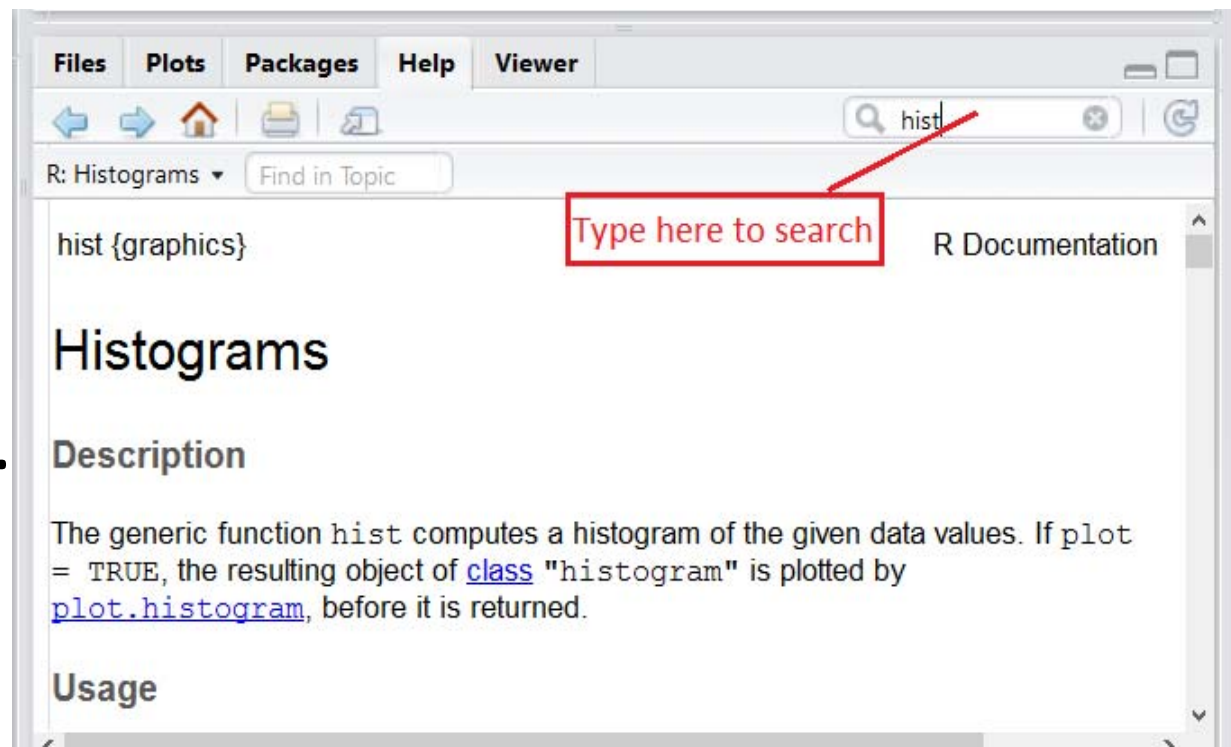
Window 4 : Output window

Help:

Various types of help can be asked.

E.g., to know about histogram,
type **hist**.

Information appears.



Essentials of Data Science With R Software - 1

Probability and Statistical Inference

Introduction to R Software

:::

Lecture 4

Calculations with R as a Calculator

Shalabh

Department of Mathematics and Statistics

Indian Institute of Technology Kanpur

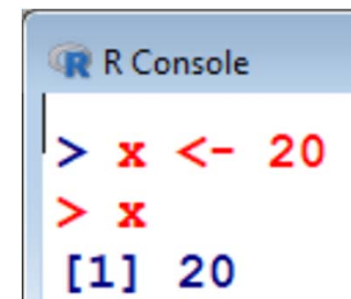
Basics

> is the prompt sign in R.

The assignment operators are the left arrow with dash <-

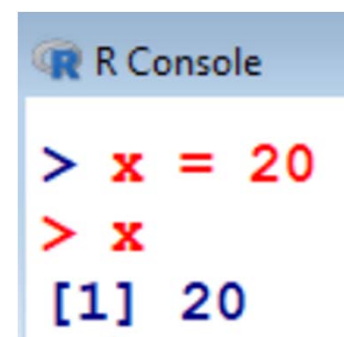
and equal sign =.

> x <- 20 assigns the value 20 to x.



```
R Console
> x <- 20
> x
[1] 20
```

> x = 20 assigns the value 20 to x.



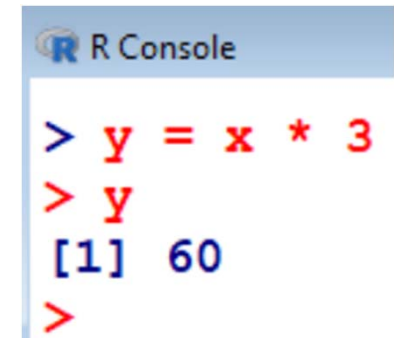
```
R Console
> x = 20
> x
[1] 20
```

Initially only <- was available in R.

Basics

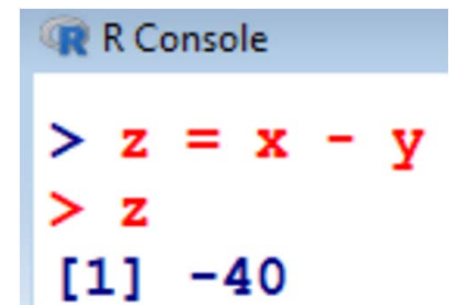
> **x = 20** assigns the value 20 to **x**.

> **y = 3 * x** assigns the value **3 * x** to **y**.



```
R Console
> y = x * 3
> y
[1] 60
>
```

> **z = x - y** assigns the value **x - y** to **z**.



```
R Console
> z = x - y
> z
[1] -40
```

Basics

The command `c(1,2,3,4)` combines the numbers 1,2,3 and 4 to a vector.

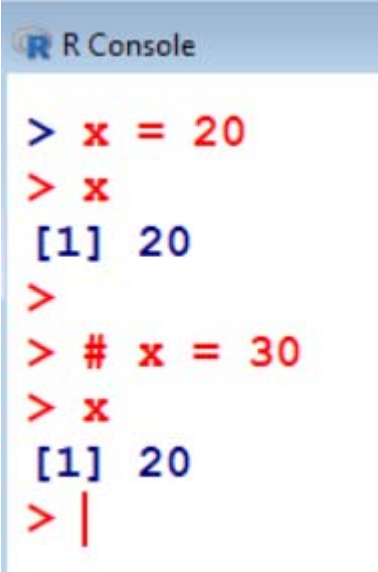
Basics

: The character # marks the beginning of a comment.

All characters until the end of the line are ignored.

> **# mu** is the mean

> **# x = 20** is treated as comment only



```
R Console
> x = 20
> x
[1] 20
>
> # x = 30
> x
[1] 20
> |
```


Basics

Capital and small letters are different.

`> X = 20` and `> x = 20` are different

R Console

```
> X = 20
```

```
> X
```

```
[1] 20
```

R Console

```
> x=20
```

```
> x
```

```
[1] 20
```

```
>
```

```
> X
```

```
Error: object 'X' not found
```

```
>
```

```
> x=10
```

```
> x
```

```
[1] 10
```

```
>
```

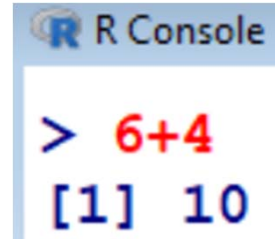
```
> x
```

```
[1] 20
```

R as a calculator

Addition

```
> 6+4          # Command  
[1] 10         # Output
```

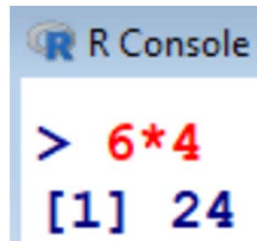


R Console

```
> 6+4  
[1] 10
```

Multiplication

```
> 6*4          # Command  
[1] 24         # Output
```



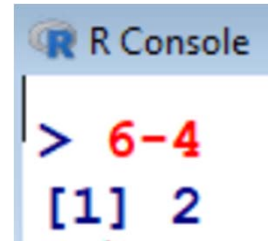
R Console

```
> 6*4  
[1] 24
```

R as a calculator

Subtraction

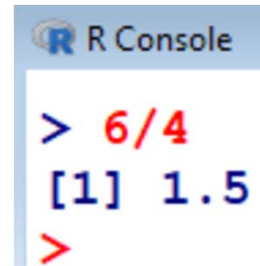
```
> 6-4      # Command  
[1] 2      # Output
```

A screenshot of the R Console window showing the command > 6-4 and the output [1] 2. The prompt > is red, and the numbers 6 and 4 are red, while the output [1] 2 is blue.

```
R Console  
> 6-4  
[1] 2
```

Division

```
> 6/4      # Command  
[1] 1.5     # Output
```

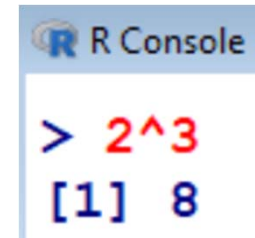
A screenshot of the R Console window showing the command > 6/4 and the output [1] 1.5. The prompt > is red, and the numbers 6 and 4 are red, while the output [1] 1.5 is blue. A red > prompt is visible on the line below.

```
R Console  
> 6/4  
[1] 1.5  
>
```

R as a calculator

Power

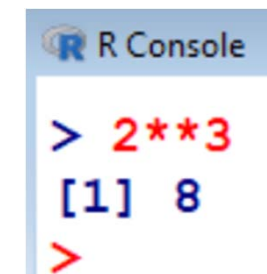
```
> 2^3          # Command  
[1] 8          # Output
```



R Console

```
> 2^3  
[1] 8
```

```
> 2**3         # Command  
[1] 8           # Output
```



R Console

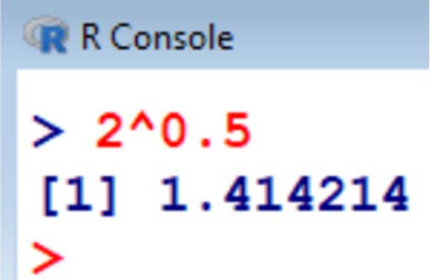
```
> 2**3  
[1] 8  
>
```

2^3

R as a calculator

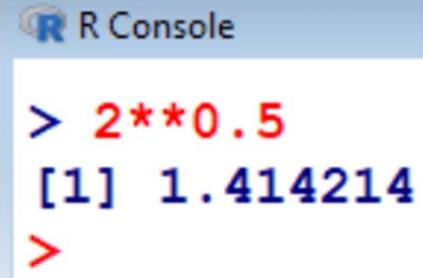
Power

```
> 2^0.5      # Command  
[1] 1.732051  # Output
```



A screenshot of the R Console window. The title bar says "R Console". The prompt ">" is followed by the command "2^0.5" in red. The output "[1] 1.414214" is shown in blue. A red ">" prompt is at the bottom.

```
> 2**0.5     # Command  
[1] 1.732051  # Output
```



A screenshot of the R Console window. The title bar says "R Console". The prompt ">" is followed by the command "2**0.5" in red. The output "[1] 1.414214" is shown in blue. A red ">" prompt is at the bottom.

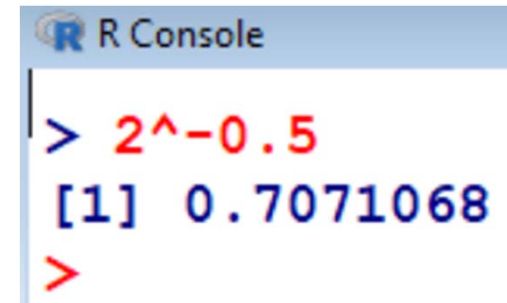
$2^{1/2}$

R as a calculator

Power

```
> 2^-0.5      # Command  
[1] 0.5773503  # Output
```

$$2^{-1/2}$$

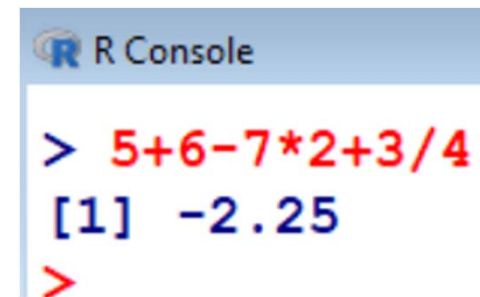


```
R Console  
> 2^-0.5  
[1] 0.7071068  
>
```

Multiple operators (BODMAS)

Bracket, Of, Division, Multiplication, Addition, and Subtraction

```
> 5+6-7*2+3/4  # Command  
[1] -2.25       # Output
```



```
R Console  
> 5+6-7*2+3/4  
[1] -2.25  
>
```

Essentials of Data Science With R Software - 1

Probability and Statistical Inference

Introduction to R Software

:::

Lecture 5

Calculations with Data Vectors

Shalabh

Department of Mathematics and Statistics

Indian Institute of Technology Kanpur

Calculations with Data Vectors

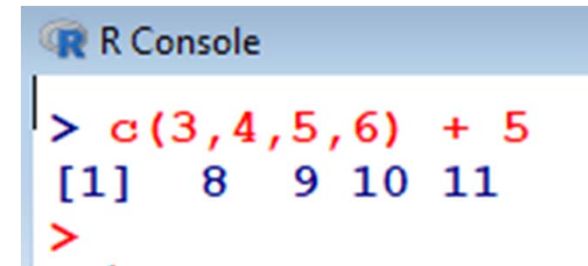
How R behaves with data vectors?

**What happens when a scalar is added/subtracted/multiplied/
divided in a data vector?**

Addition in Data Vectors: $x + y$

```
> c(3,4,5,6) + 5  
[1] 8  9 10 11
```

3+5, 4+5, 5+5, 6+5

A screenshot of an R console window. The title bar says "R Console". The prompt ">" is followed by the command "c(3,4,5,6) + 5" in red. The output "[1] 8 9 10 11" is shown in blue. The prompt ">" is followed by a period "." in red.

```
R Console  
> c(3,4,5,6) + 5  
[1] 8 9 10 11  
> .
```

Addition in Data Vectors: $x + y$

```
> c(3, 4, 5, 6) + c(-3, -4, -5, 7)
```

```
[1] 0 0 0 13
```

$3+(-3), 4+(-4), 5+(-5), 6+7$

R Console

```
> > c(3,4,5,6) + c(-3,-4, -5, 7)
```

```
[1] 0 0 0 13
```


```
>
```

Addition in Data Vectors: $x + y$

```
> c(3,4,5,6) + c(7,8)
```

```
[1] 10 12 12 14
```

3+7, 4+8, 5+7, 6+8

 R Console

```
> c(3,4,5,6) + c(7,8)
```

```
[1] 10 12 12 14
```

```
>
```

Addition in Data Vectors: $x + y$

```
> c(3,4,5,6) + c(7,8,9) # Warning message
```

```
[1] 10 12 14 13
```

Warning message:

```
In: c(3, 4, 5, 6) + c(7, 8, 9)
```

```
longer object length is not a multiple of  
shorter object length
```

3+7, 4+8, 5+9, 6+7

Addition in Data Vectors: $x + y$

R Console

```
> c(3,4,5,6) + c(7,8,9)
[1] 10 12 14 13
Warning message:
In c(3, 4, 5, 6) + c(7, 8, 9) :
  longer object length is not a multiple of shorter object length
>
```

Subtraction in Data Vectors: $x - y$

```
> c(3,4,5,6) - 5
```

```
[1] -2  -1  0  1
```

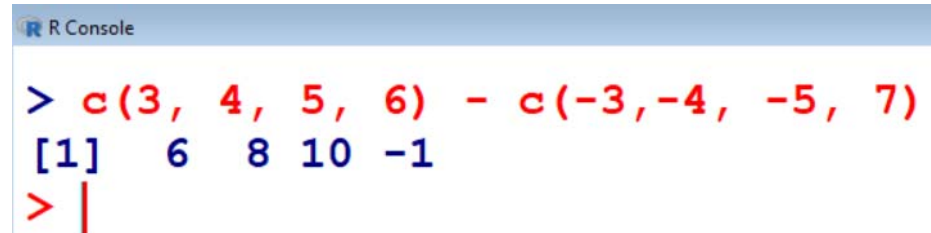
3-5, 4-5, 5-5, 6-5

Subtraction in Data Vectors: $x - y$

```
> c(3, 4, 5, 6) - c(-3, -4, -5, 7)
```

```
[1] 6 8 10 -1
```

$3 - (-3), 4 - (-4), 5 - (-5), 6 - 7$

A screenshot of an R console window. The title bar says "R Console". The prompt ">" is followed by the command "c(3, 4, 5, 6) - c(-3, -4, -5, 7)". The output is "[1] 6 8 10 -1". Below the output, the prompt ">" is followed by a vertical bar "|".

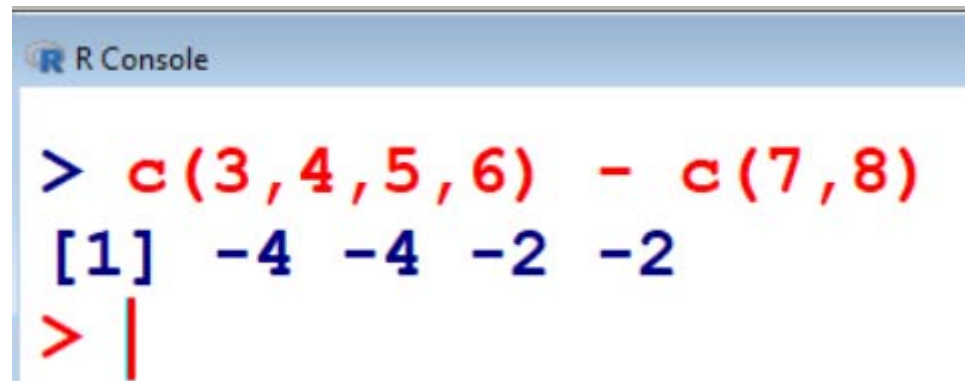
```
R Console  
> c(3, 4, 5, 6) - c(-3, -4, -5, 7)  
[1] 6 8 10 -1  
> |
```

Subtraction in Data Vectors: $x - y$

```
> c(3,4,5,6) - c(7,8)
```

```
[1] -4 -4 -2 -2
```

3-7, 4-8, 5-7, 6-8

A screenshot of an R console window. The title bar says "R Console". The prompt ">" is followed by the command "c(3,4,5,6) - c(7,8)" in red text. Below it, the output "[1] -4 -4 -2 -2" is shown in blue text. A new prompt ">" with a vertical cursor is shown on the next line.

```
> c(3,4,5,6) - c(7,8)
[1] -4 -4 -2 -2
> |
```


Subtraction in Data Vectors: $x - y$

```
> c(3,4,5,6) - c(7,8,9) # Warning message
```

```
[1] -4 -4 -4 -1
```

Warning message:

```
In c(3, 4, 5, 6) - c(7, 8, 9) :
```

longer object length is not a multiple of
shorter object length

3-7, 4-8, 5-9, 6-7

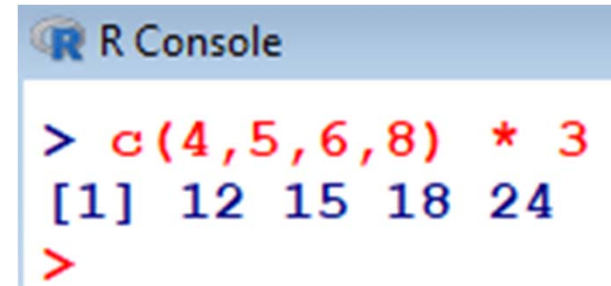
```
R Console
> c(3,4,5,6) - c(7,8,9)
[1] -4 -4 -4 -1
Warning message:
In c(3, 4, 5, 6) - c(7, 8, 9) :
  longer object length is not a multiple of shorter object length
> |
```

Multiplication in Data Vectors: $x * y$

```
> c(4,5,6,8) * 3
```

```
[1] 12 15 18 24
```

$4 \times 3, 5 \times 3, 6 \times 3, 8 \times 3$

A screenshot of an R console window. The title bar is light blue and contains the R logo and the text "R Console". The console area has a white background and shows the command "> c(4,5,6,8) * 3" in red text, followed by the output "[1] 12 15 18 24" in blue text. A red prompt character ">" is visible on the next line.

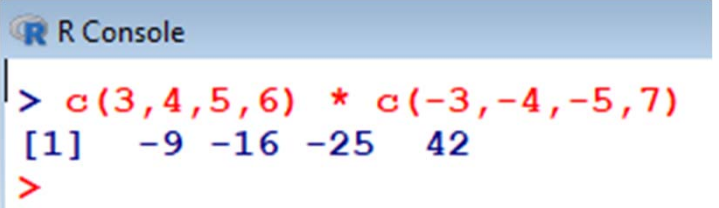
```
> c(4,5,6,8) * 3
[1] 12 15 18 24
>
```

Multiplication in Data Vectors: $x * y$

```
> c(3,4,5,6) * c(-3,-4,-5,7)
```

```
[1] -9 -16 -25 42
```

$3 \times (-3), 4 \times (-4), 5 \times (-5), 6 \times 7$



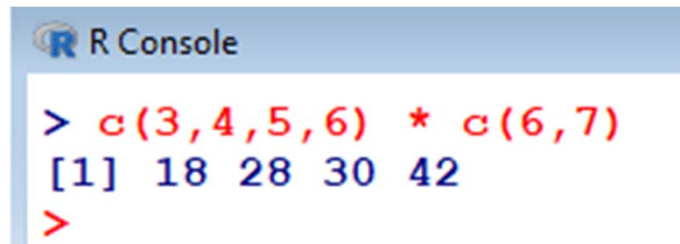
```
R Console  
> c(3,4,5,6) * c(-3,-4,-5,7)  
[1] -9 -16 -25 42  
>
```

Multiplication in Data Vectors: $x * y$

```
> c(3,4,5,6) * c(6,7) # !!! ATTENTION
```

```
[1] 18 28 30 42
```

$3 \times 6, 4 \times 7, 5 \times 6, 6 \times 7$

A screenshot of an R console window. The title bar says "R Console". The prompt ">" is followed by the command "c(3,4,5,6) * c(6,7)" in red text. Below it, the output "[1] 18 28 30 42" is shown in blue text. The prompt ">" is shown again at the bottom in red text.

```
> c(3,4,5,6) * c(6,7)
[1] 18 28 30 42
>
```

Multiplication in Data Vectors: $x * y$

```
> c(3,4,5,6) * c(7,8,9) # Warning message
```

```
[1] 21 32 45 42
```

Warning message:

In $c(3, 4, 5, 6) * c(7, 8, 9)$

longer object length

is not a multiple of shorter object length

3 x 7, 4 x 8, 5 x 9, 6 x 7

```
R Console
> c(3,4,5,6) * c(7,8,9)
[1] 21 32 45 42
Warning message:
In c(3, 4, 5, 6) * c(7, 8, 9) :
  longer object length is not a multiple of shorter object length
> |
```

Division in Data Vectors: x / y

```
> c(2,4,6,8) / 2
```

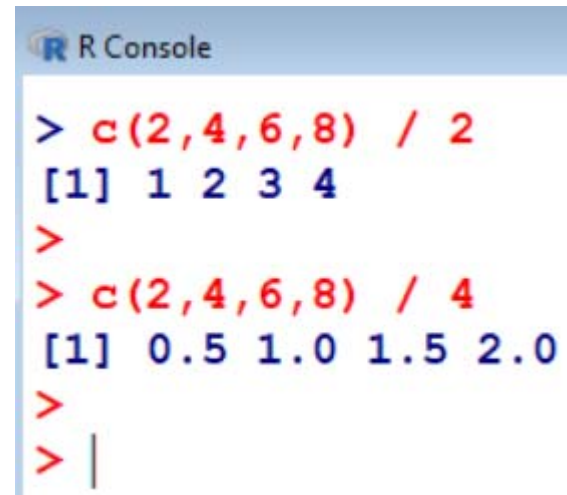
```
[1] 1 2 3 4
```

$2 / 2, 4 / 2, 6 / 2, 8 / 2$

```
> c(2,4,6,8) / 4
```

```
[1] 0.5 1.0 1.5 2.0
```

$2 / 4, 4 / 4, 6 / 4, 8 / 4$



R Console

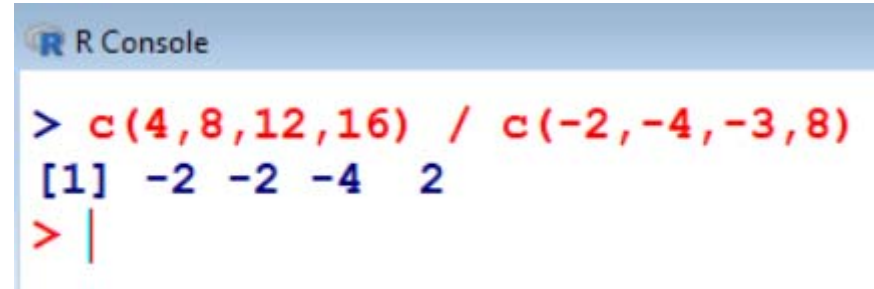
```
> c(2,4,6,8) / 2  
[1] 1 2 3 4  
>  
> c(2,4,6,8) / 4  
[1] 0.5 1.0 1.5 2.0  
>  
> |
```

Division in Data Vectors: x / y

```
> c(4,8,12,16) / c(-2,-4,-3,8)
```

```
[1] -2 -2 -4 2
```

$4 / (-2), 8 / (-4), 12 / (-3), 16 / 8$



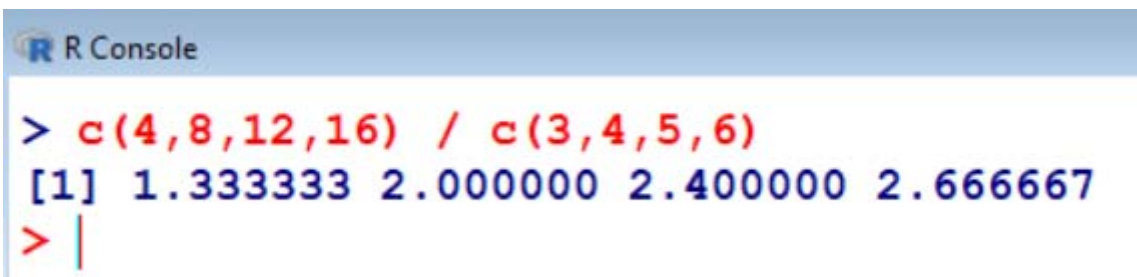
```
R Console  
> c(4,8,12,16) / c(-2,-4,-3,8)  
[1] -2 -2 -4 2  
> |
```

Division in Data Vectors: x / y

```
> c(4,8,12,16) / c(3,4,5,6)
```

```
[1] 1.333333 2.000000 2.400000 2.666667
```

$4 / 3, 8 / 4, 12 / 5, 16 / 6$

A screenshot of an R console window. The title bar is light blue and says "R Console". The console area has a white background. It shows the same R command and output as the main slide: a red prompt character followed by the command `c(4,8,12,16) / c(3,4,5,6)`, then a blue prompt character followed by the output `[1] 1.333333 2.000000 2.400000 2.666667`. A red prompt character is on the next line, followed by a vertical cursor bar.

```
R Console
> c(4,8,12,16) / c(3,4,5,6)
[1] 1.333333 2.000000 2.400000 2.666667
> |
```


Division in Data Vectors: x / y

```
> c(4,8,12,16) / c(2,4,3) # Warning message  
[1] 2 2 4 8
```

Warning message:

In $c(4, 8, 12, 16)/c(2, 4, 3)$:

longer object length is not a multiple of
shorter object length

$4 / 2, 8 / 4, 12 / 3, 16 / 2$

R Console

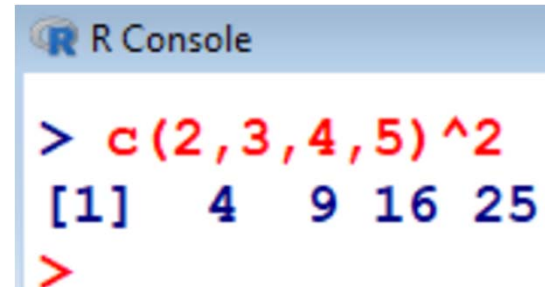
```
> c(3,4,5,6) * c(7,8,9)  
[1] 21 32 45 42  
Warning message:  
In c(3, 4, 5, 6) * c(7, 8, 9) :  
longer object length is not a multiple of shorter object length  
> |
```

Power Operators in Data Vectors:

```
> c(2,3,4,5)^2  
[1] 4 9 16 25
```

command: application to a vector
output

$2^2, 3^2, 4^2, 5^2$

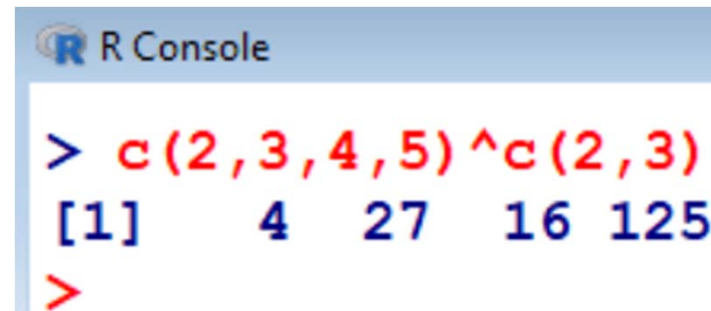


```
R Console  
> c(2,3,4,5)^2  
[1] 4 9 16 25  
>
```

Power Operators in Data Vectors:

```
> c(2,3,4,5)^c(2,3) # !!ATTENTION! Observe the  
[1] 4 27 16 125      # operation  
# output
```

$2^2, 3^3, 4^2, 5^3$

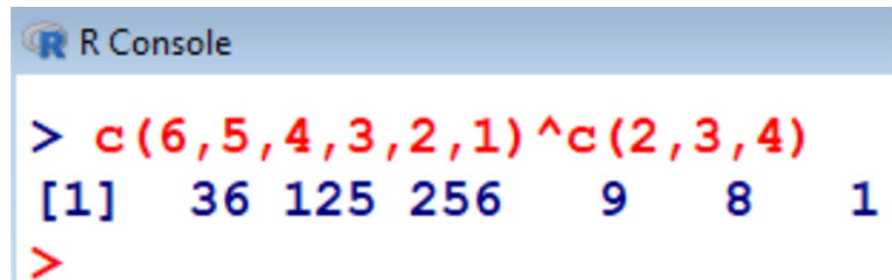


```
R Console  
> c(2,3,4,5)^c(2,3)  
[1] 4 27 16 125  
>
```

Power Operators in Data Vectors:

```
> c(6,5,4,3,2,1)^c(2,3,4) # command: application  
                             to a vector with vector  
[1] 36 125 256 9 8 1 # output
```

$6^2, 5^3, 4^4, 3^2, 2^3, 1^4$



```
R Console  
> c(6,5,4,3,2,1)^c(2,3,4)  
[1] 36 125 256 9 8 1  
>
```

Power Operators in Data Vectors:

```
> c(6,5,4,3)^c(3,4,5)      # Warning message  
[1] 216  625 1024   27      # output
```

Warning message:

In `c(6,5,4,3)^c(3,4,5)` :longer object length is not a multiple of shorter object length

$6^3, 5^4, 4^5, 3^3$

R Console

```
> c(6,5,4,3)^c(3,4,5)  
[1] 216  625 1024   27  
Warning message:  
In c(6, 5, 4, 3)^c(3, 4, 5) :  
  longer object length is not a multiple of shorter object length  
> |
```

Essentials of Data Science With R Software - 1

Probability and Statistical Inference

Introduction to R Software

:::

Lecture 6

Built-in Commands and Bivariate Plots

Shalabh

Department of Mathematics and Statistics

Indian Institute of Technology Kanpur

Built in commands

Some commands are readily available in R to compute the mathematical functions.

How to use them and utilize them in computing various quantities?

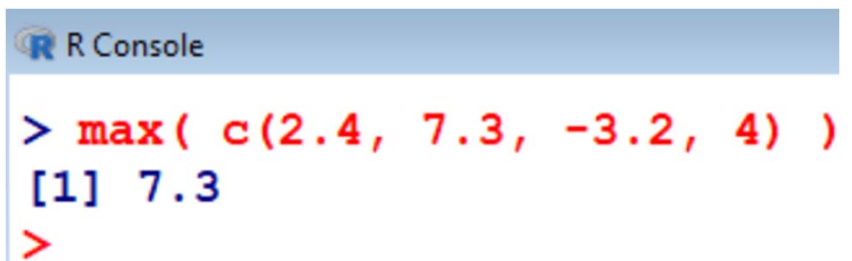
Maximum

```
> max(2.4, 7.3, -3.2, 4)
[1] 7.3
```

A screenshot of an R console window with a blue header bar containing the R logo and the text "R Console". The console shows the command `> max(2.4, 7.3, -3.2, 4)` in red text, followed by the output `[1] 7.3` in blue text, and a red prompt character `>` on the next line.

```
> max(2.4, 7.3, -3.2, 4)
[1] 7.3
>
```

```
> max( c(2.4, 7.3, -3.2, 4) )
[1] 7.3
```

A screenshot of an R console window with a blue header bar containing the R logo and the text "R Console". The console shows the command `> max(c(2.4, 7.3, -3.2, 4))` in red text, followed by the output `[1] 7.3` in blue text, and a red prompt character `>` on the next line.

```
> max( c(2.4, 7.3, -3.2, 4) )
[1] 7.3
>
```

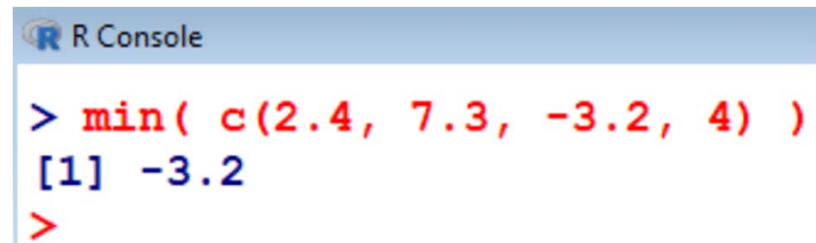

Minimum

```
> min(2.4, 7.3, -3.2, 4)
[1] -3.2
```

A screenshot of the R Console window. The title bar is light blue and contains the R logo and the text "R Console". The console area has a white background and shows the command "> min(2.4, 7.3, -3.2, 4)" in red text, followed by the output "[1] -3.2" in blue text. A red prompt character ">" is visible on the next line.

```
> min(2.4, 7.3, -3.2, 4)
[1] -3.2
>
```

```
> min( c(2.4, 7.3, -3.2, 4) )
[1] -3.2
```

A screenshot of the R Console window. The title bar is light blue and contains the R logo and the text "R Console". The console area has a white background and shows the command "> min(c(2.4, 7.3, -3.2, 4))" in red text, followed by the output "[1] -3.2" in blue text. A red prompt character ">" is visible on the next line.

```
> min( c(2.4, 7.3, -3.2, 4) )
[1] -3.2
>
```

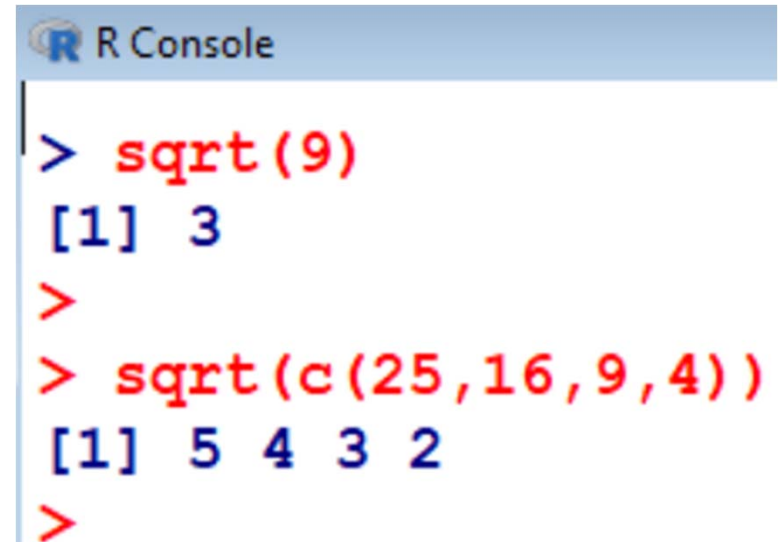
Overview Over Other Functions

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

Examples

```
> sqrt(9)
[1] 3
```

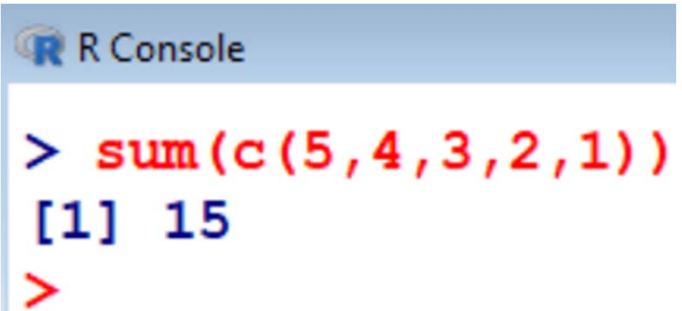
```
> sqrt(c(25,16,9,4))
[1] 5 4 3 2
```

A screenshot of an R console window. The title bar is blue with the R logo and the text "R Console". The console shows two commands and their outputs. The first command is "> sqrt(9)" in red, followed by the output "[1] 3" in blue. The second command is "> sqrt(c(25,16,9,4))" in red, followed by the output "[1] 5 4 3 2" in blue. The prompt ">" is shown in red at the end of each line.

```
R Console
> sqrt(9)
[1] 3
>
> sqrt(c(25,16,9,4))
[1] 5 4 3 2
>
```

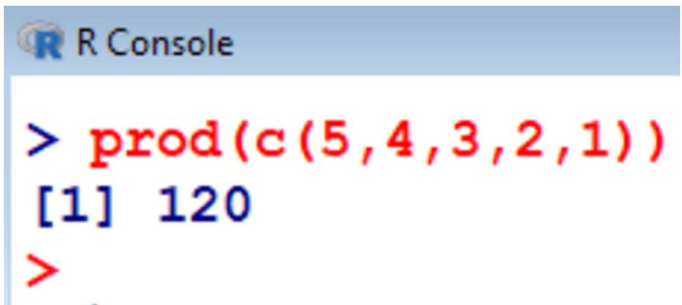
Examples

```
> sum(c(5,4,3,2,1))  
[1] 15
```

A screenshot of an R console window. The title bar is light blue and contains the R logo and the text "R Console". The console area is white and shows the command `> sum(c(5,4,3,2,1))` in red text, followed by the output `[1] 15` in blue text. A red prompt character `>` is on the next line.

```
> sum(c(5,4,3,2,1))  
[1] 15  
>
```

```
> prod(c(5,4,3,2,1))  
[1] 120
```

A screenshot of an R console window. The title bar is light blue and contains the R logo and the text "R Console". The console area is white and shows the command `> prod(c(5,4,3,2,1))` in red text, followed by the output `[1] 120` in blue text. A red prompt character `>` is on the next line.

```
> prod(c(5,4,3,2,1))  
[1] 120  
>
```

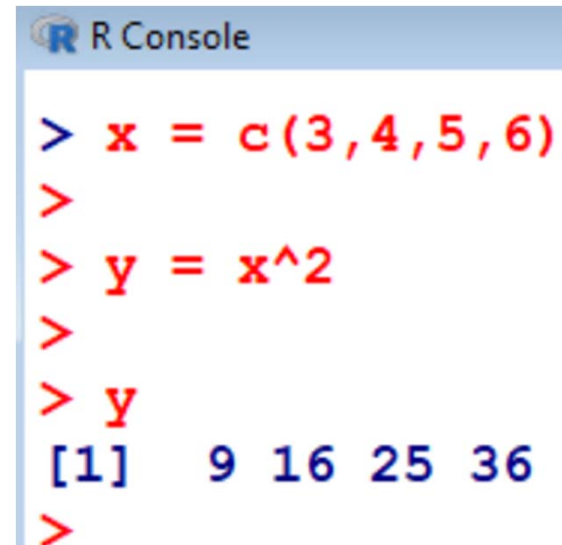
Assignments

An assignment can also be used to save values in variables:

```
> x = c(3,4,5,6)
```

```
> y = x^2
```

```
> y  
[1] 9 16 25 36
```

A screenshot of the R Console window. The title bar says "R Console". The console shows the following commands and output:

```
> x = c(3,4,5,6)  
>  
> y = x^2  
>  
> y  
[1] 9 16 25 36  
>
```

The commands and their corresponding outputs are shown in red text. The prompt character is a blue greater-than sign (>).

Bivariate plots:

Provide first hand visual information about the nature and degree of relationship between two variables.

Relationship can be linear or nonlinear.

We discuss several types of plots through examples.

Bivariate plots: Scatter plot

Plot command:

x, y: Two data vectors

`plot(x, y)`

`plot(x, y, type)`

type	
"p" for <u>p</u> oints	"l" for <u>l</u> ines
"b" for <u>b</u> oth	"c" for the lines part alone of "b"
"o" for both ' <u>o</u> verplotted'	"s" for stair <u>s</u> teps.
"h" for ' <u>h</u> istogram' like (or 'high-density') vertical lines	

Bivariate plots: Scatter plot

Plot command:

x, y: Two data vectors

```
plot(x, y)
```

```
plot(x, y, type)
```

Get more details from help: `help("type")`

Other options:

main an overall title for the plot.

suba sub title for the plot.

xlaba title for the x axis.

ylaba title for the y axis.

aspthe y/x aspect ratio.

Bivariate plots: Example

Number of marks obtained by students depend upon the number of hours of study.

Data on marks out of 500 maximum marks and number of hours per week for 20 students are collected as follows:

Marks out of 500 maximum marks

```
marks <- c(337,316,334,327,340,360, 374,330,352,  
353,370,380,384,398,413,428,430,438,439,450)
```

Number of hours per week

```
hours <- c(23,25,25,26,27,28,30,26,29,32,33,34,  
35,38,39,42,43,44, 45,45.5)
```

Bivariate plots: Scatter plot

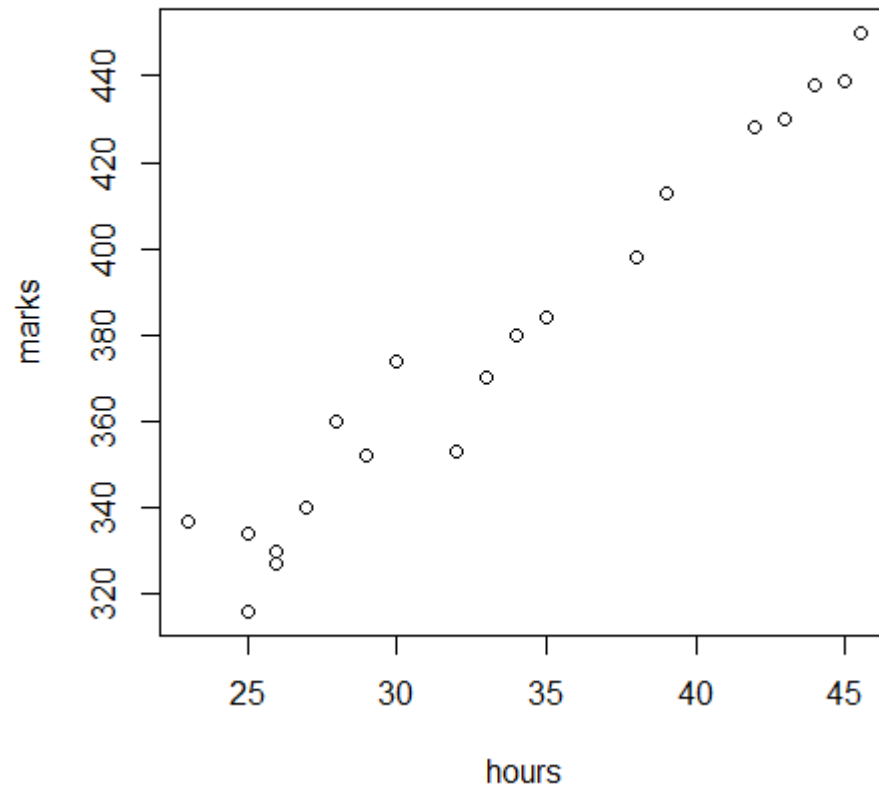
`plot` command:

`x, y`: Two data vectors

Various type of plots are possible to draw.

`plot(x, y)`

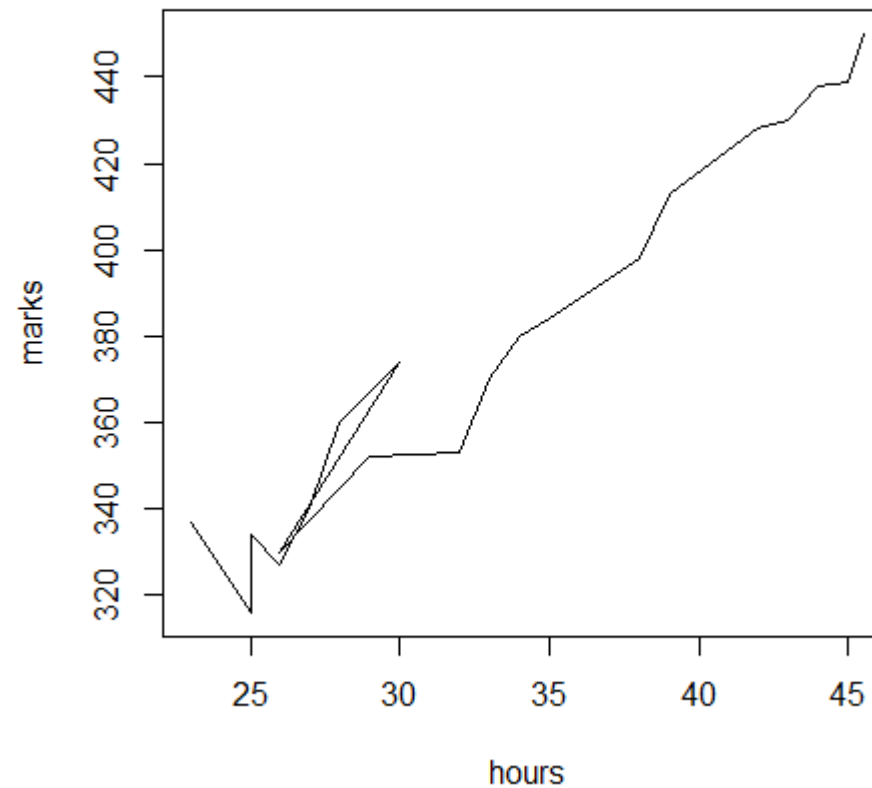
`plot(hours, marks)`



Bivariate plots: Scatter plot

```
plot(hours, marks, "l")
```

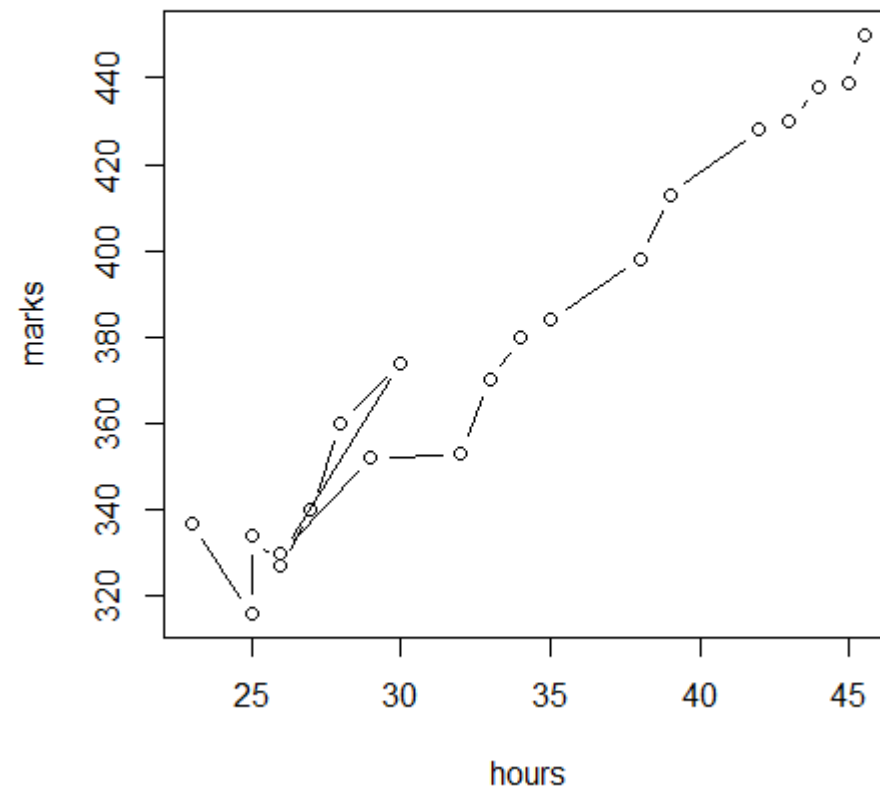
“l” for lines,



Bivariate plots: Scatter plot

```
plot(hours, marks, "b")
```

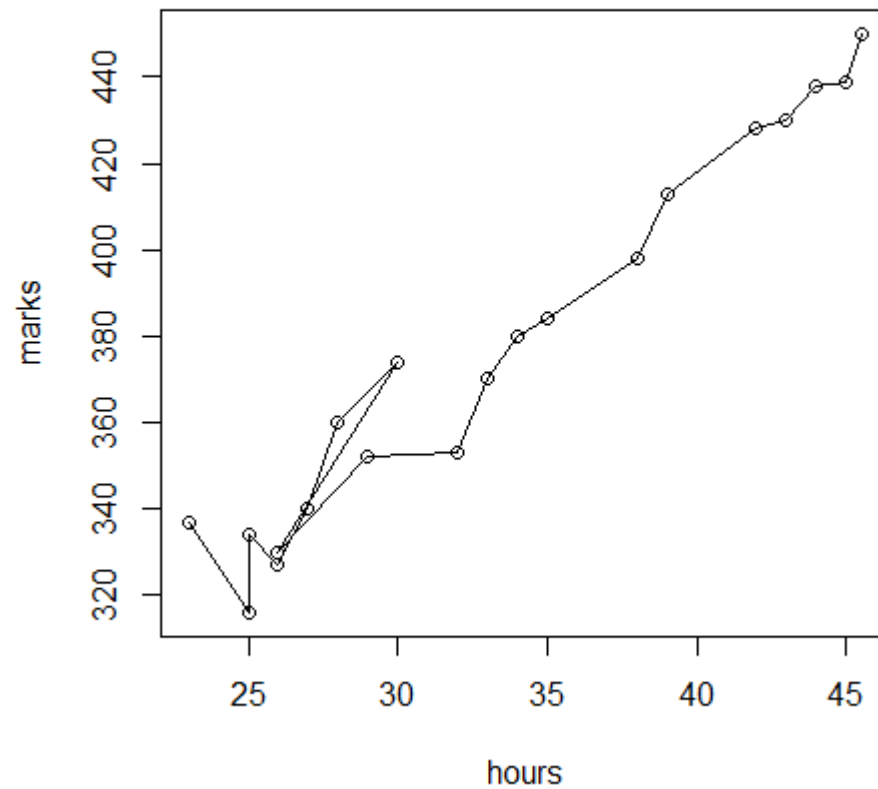
"b" for both – line and point



Bivariate plots: Scatter plot

```
plot(hours, marks, "o")
```

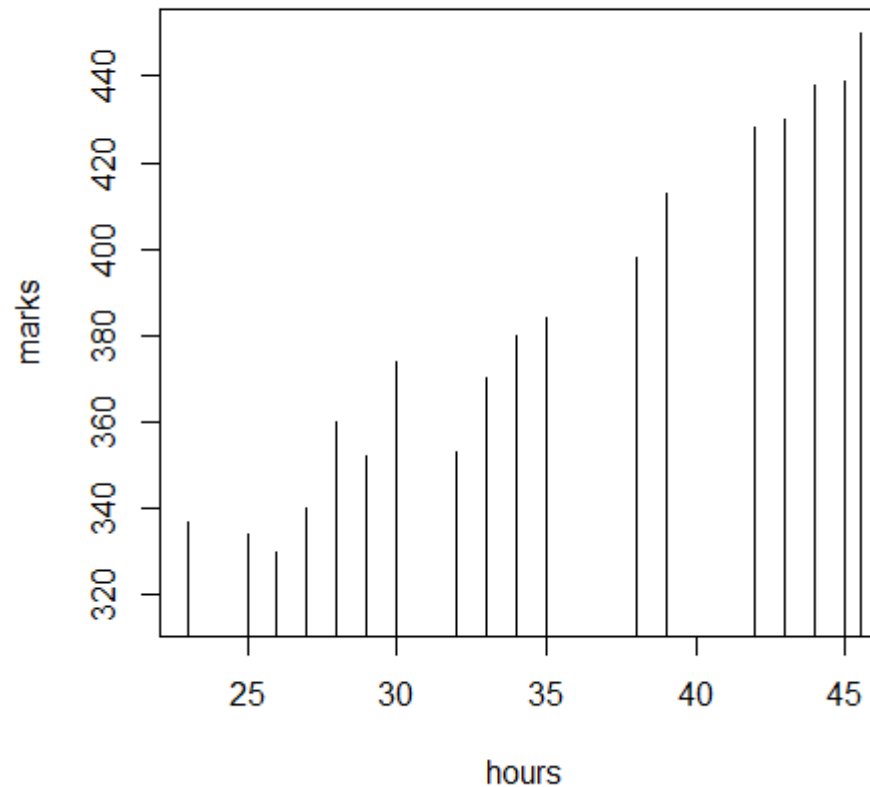
“o” for both ‘overplotted’



Bivariate plots: Scatter plot

```
plot(hours, marks, "h")
```

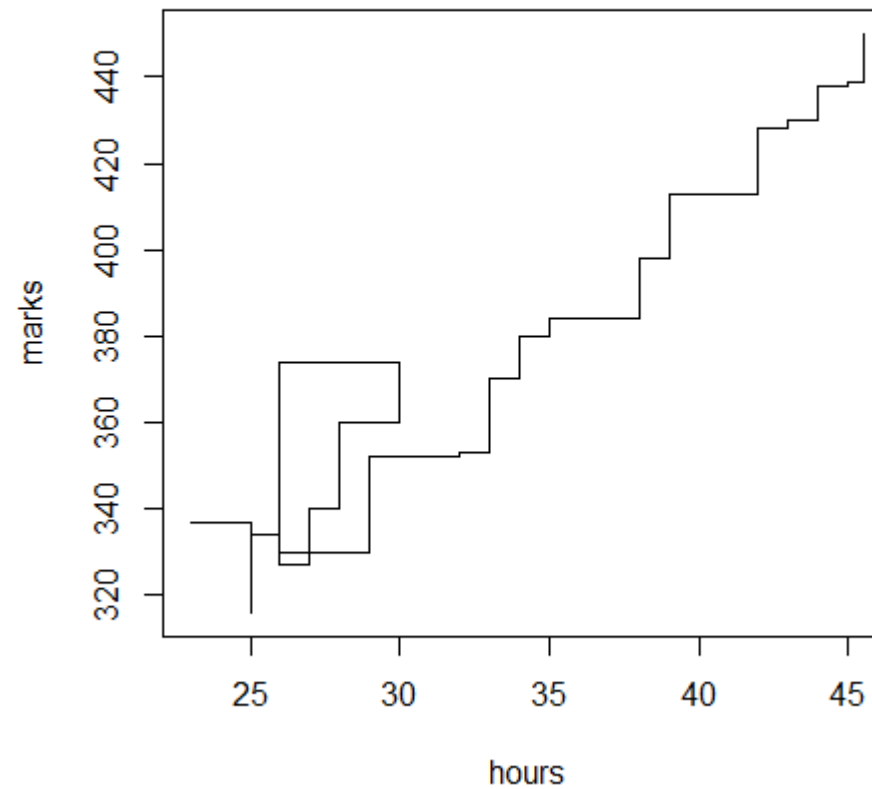
"h" for 'histogram' like (or 'high-density') vertical lines



Bivariate plots: Scatter plot

```
plot(hours, marks, "s")
```

“s” for stair steps.



Essentials of Data Science With R Software - 1

Probability and Statistical Inference

Introduction to R Software

:::

Lecture 7

Logical Operators and Selection of Sample

Shalabh

Department of Mathematics and Statistics

Indian Institute of Technology Kanpur

Logical Operators and Comparisons:

- **The cities and households are categorized in the data as 1, 2,...**
- **We want to find the mean of income those households which are in the cities coded as 1.**
- **We want to find the mean of income of the households in cities coded as 1 having household size more than 3.**
- **Less than, more than and not equal to are the logical operations, not mathematical operations.**

Logical Operators and Comparisons

The following table shows the operations and functions for logical comparisons (True or False).

TRUE and **FALSE** are reserved words denoting logical constants.

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

Logical Operators and Comparisons

TRUE and FALSE are reserved words denoting logical constants

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

Examples:

```
> 8 > 7  
[1] TRUE
```

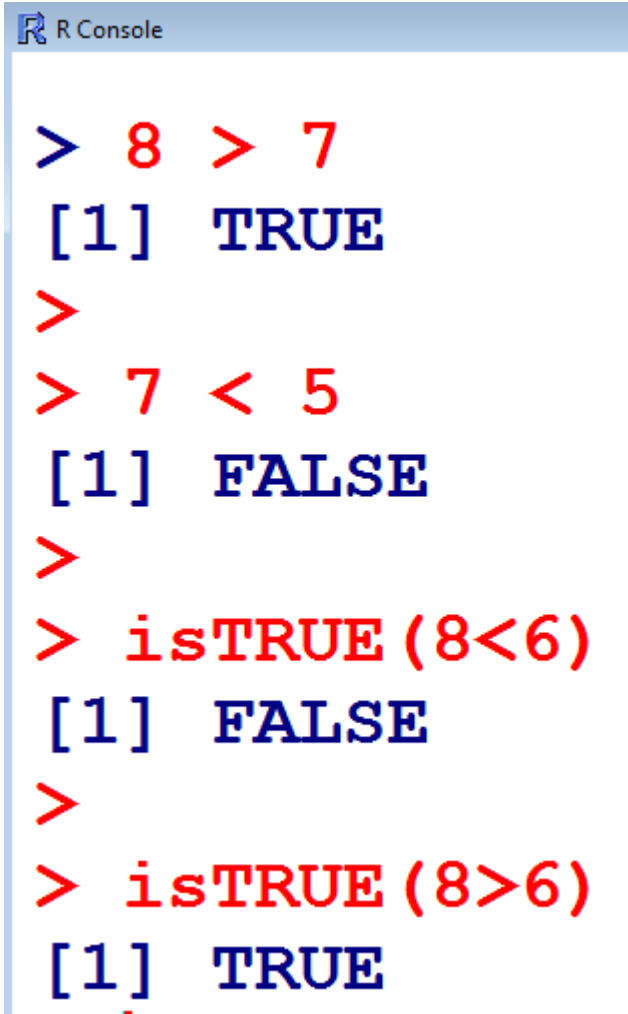
```
> 7 < 5  
[1] FALSE
```

Is 8 less than 6?

```
> isTRUE(8<6)  
[1] FALSE
```

Is 8 greater than 6?

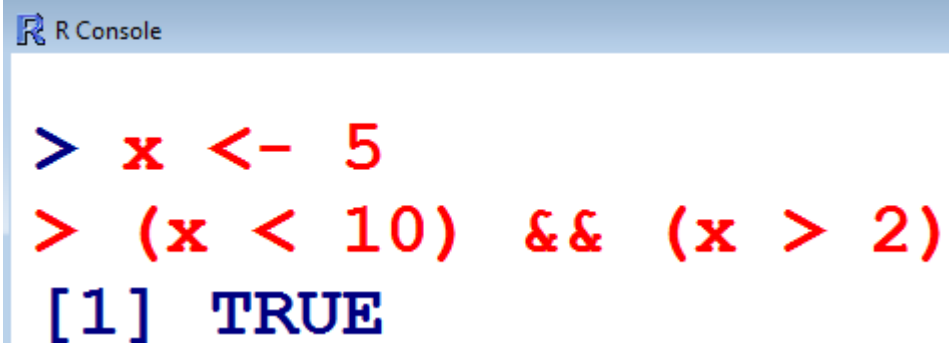
```
> isTRUE(8>6)  
[1] TRUE
```

A screenshot of an R Console window. The window has a title bar that says "R Console". The console contains several lines of R code and their corresponding output. The code is color-coded: the prompt ">" is blue, and the code itself is red. The output is blue. The code includes logical comparisons and the isTRUE() function.

```
R Console  
  
> 8 > 7  
[1] TRUE  
  
>  
> 7 < 5  
[1] FALSE  
  
>  
> isTRUE(8<6)  
[1] FALSE  
  
>  
> isTRUE(8>6)  
[1] TRUE
```

Examples:

```
> x <- 5  
> (x < 10) && (x > 2)    # && means AND  
[1] TRUE
```



The image shows a screenshot of an R Console window. The window has a title bar that says "R Console". Inside the console, the following R code is entered and executed:

```
> x <- 5  
> (x < 10) && (x > 2)  
[1] TRUE
```

Examples:

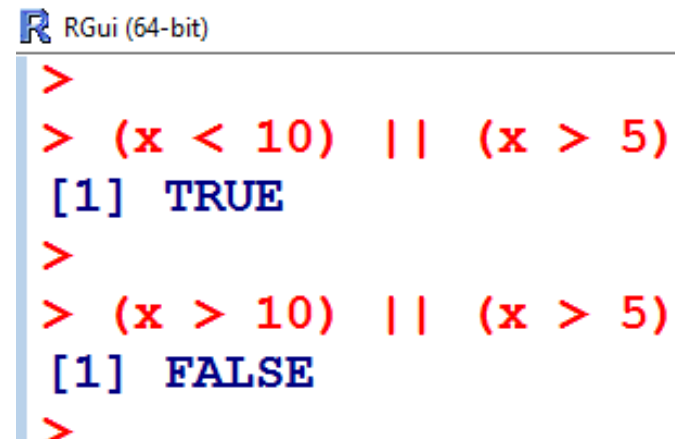
```
> x <- 5
```

Is **x** less than 10 or **x** is greater than 5 ?

```
> (x < 10) || (x > 5)      # || means OR  
[1] TRUE
```

Is **x** greater than 10 or **x** is greater than 5 ?

```
> (x > 10) || (x > 5)  
[1] FALSE
```



```
RGui (64-bit)  
>  
> (x < 10) || (x > 5)  
[1] TRUE  
>  
> (x > 10) || (x > 5)  
[1] FALSE  
>
```

Examples:

```
> x = 10
```

```
> y = 20
```

Is **x** equal to 10 and is **y** equal to 20?

```
> (x == 10) & (y == 20)
```

```
[1] TRUE
```

== means exactly
equal to

Is **x** equal to 10 and is **y** equal to 2?

```
> (x == 10) & (y == 2)
```

```
[1] FALSE
```

R Console

```
> x = 10
```

```
> y = 20
```

```
>
```

```
> (x == 10) & (y == 20)
```

```
[1] TRUE
```

```
>
```

```
> (x == 10) & (y == 2)
```

```
[1] FALSE
```

Simple Random Sampling:

Simple random sampling (SRS) is a method of selection of a sample comprising of n number of sampling units from the population having N number of units such that every sampling unit has an equal chance of being chosen.

Simple Random Sampling Without and With Replacement:

SRSWOR

The sampling units are chosen without replacement in the sense that the units once chosen are not placed back in the population .

SRSWR

The sampling units are chosen with replacement in the sense that the chosen units are placed back in the population.

Drawing a Simple Random Sample Without Replacement (SRSWOR)

`sample` takes a sample of the specified size from the elements of `x` using either with or without replacement.

Usage

```
sample(x, size, replace = FALSE)
```

Arguments

`x` Either a vector of one or more elements from which to choose, or a positive integer.

`size` a non-negative integer giving the number of items to choose.

`replace` Should sampling be with replacement?

Drawing a Simple Random Sample Without Replacement (SRSWOR)

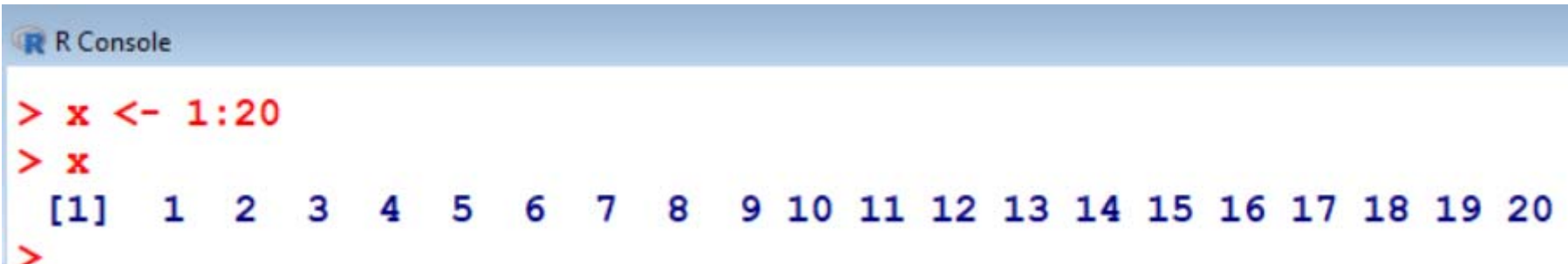
First we define a population units containing the numbers 1 to 20.

This can be defined by a sequence as **x**.

```
> x <- 1:20
```

```
> x
```

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

A screenshot of an R console window. The title bar says "R Console". The console shows the same R code and output as the previous block: the command to create vector x, the command to print x, and the resulting sequence of numbers from 1 to 20.

```
> x <- 1:20  
> x  
[1] 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20  
>
```

Drawing a Simple Random Sample Without Replacement (SRSWOR)

Let us draw the sample of size 5 from population **x** by SRSWOR .

This is controlled by the statement **replace = FALSE** inside the argument.

```
> x  
[1] 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17  
18 19 20
```

SRSWOR command

```
sample(x, size=5, replace = FALSE)
```

Drawing a Simple Random Sample Without Replacement (SRSWOR)

```
> sample(x, size=5, replace = FALSE)
```

```
[1] 15  1 10 11  5
```

```
> sample(x, size=5, replace = FALSE)
```

```
[1] 13  9 10 17 20
```

```
> sample(x, size=5, replace = FALSE)
```

```
[1] 11  8  5 12 13
```

Drawing a Simple Random Sample Without Replacement (SRSWOR)

```
R Console  
  
> sample(x, size=5, replace = FALSE)  
[1] 15  1 10 11  5  
>  
> sample(x, size=5, replace = FALSE)  
[1] 13  9 10 17 20  
>  
> sample(x, size=5, replace = FALSE)  
[1] 11  8  5 12 13  
> |
```

Drawing a Simple Random Sample With Replacement (SRSWR)

Let us draw the sample of size 10 from population **x** by SRSWR .

This is controlled by the statement **replace = TRUE** inside the argument.

```
> x  
[1] 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17  
18 19 20
```

SRSWR Command

```
sample(x, size=10, replace = TRUE)
```

Drawing a Simple Random Sample With Replacement (SRSWR)

SRSWR

```
> sample(x, size=10, replace = TRUE)
```

```
[1]  4 17  6  3 20 14 13  2 15  2
```

Value 2 is repeated.

```
> sample(x, size=10, replace = TRUE)
```

```
[1]  5 12  7  4 18  2 12  1  3  7
```

Values 12 and 7 are repeated.

```
> sample(x, size=10, replace = TRUE)
```

```
[1] 15 11 19 10  4  3 11 17  9  3
```

Value 11 is repeated.

Drawing a Simple Random Sample With Replacement (SRSWR)

R Console

```
> x
[1] 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
> sample(x, size=10, replace = TRUE)
[1] 4 17 6 3 20 14 13 2 15 2
> sample(x, size=10, replace = TRUE)
[1] 5 12 7 4 18 2 12 1 3 7
>
> sample(x, size=10, replace = TRUE)
[1] 15 11 19 10 4 3 11 17 9 3
>
```

Drawing a Simple Random Sample With Replacement (SRSWR)

```
> x
```

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

For `sample` the default for size is the number of items inferred from the first argument, so that `sample(x)` generates a random permutation of the elements of `x` (or `1:x`).

```
> sample(x)
```

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

```
> sample(x)
```

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

Drawing a Simple Random Sample With Replacement (SRSWR)

```
R Console
> x
[1] 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
> sample(x)
[1] 19 2 1 7 12 15 4 14 13 5 10 17 6 16 18 9 20 3 8 11
> sample(x)
[1] 6 15 8 2 14 9 18 12 4 17 7 5 20 13 1 16 11 3 10 19
> |
```

Essentials of Data Science With R Software - 1

Probability and Statistical Inference

Probability Theory

:::

Lecture 8

Introduction to Probability

Shalabh

Department of Mathematics and Statistics

Indian Institute of Technology Kanpur

Introduction to Statistics:

It has become a rule rather than exception that if we want to learn and know about any phenomenon or a process, one must collect data and learn from there.

Statistics is the science of learning from data.

It is related to the collection of data and then extracting the hidden information by its descriptive analysis and drawing of conclusions.

Introduction to Statistics:

Sometimes a statistical analysis begins with a given set of data

Statistics describes, summarizes, and analyse the data.

In case, the data is not available, in such cases, the statistical design of experiment is appropriately used to generate data.

At the end of the experiment, the data is described and summarized using the tools of descriptive statistics.

Inferential Statistics:

After completing the experiment, data is described and summarized with an aim to draw a statistical conclusion using the tools of inferential statistics.

The concept of chance is considered and utilized to draw a conclusion from the data.

Some assumptions are made about the chances/probabilities of obtaining the different data values. These assumptions are referred to as probability model for the data.

Inferential Statistics:

A careful description and presentation of the data enable us to infer an appropriate probability model for a given data set which can be verified by using the additional data.

The tools of statistical inference lay the foundation of the formulation of a probability model to describe the data.

Thus an understanding of statistical inference data to make valid inferences requires knowledge of the theory of probability.

Elements of Probability:

The probability of an event is subjected to various meanings or interpretations depending upon how one says it.

For example, if a medical doctor says that “there are 70% percent chances that the patient will be cured,” gives some intuitive idea about the success of treatment.

One can also say that the doctor feels that, over the long run, in 70% percent of such ailments, the patients have recovered

Interpretation of Probability:

Probability: Measure of uncertainty

Broadly, there are two types of interpretation of probability

- 1. Frequency interpretation**
- 2. Subjective interpretation of probability.**

Frequency Interpretation of Probability:

The probability of a given outcome of an experiment indicates a “property” of that outcome.

Such a property can be determined by continual repetition of the experiment.

A popular interpretation of probability is as follows:

The probability of the outcome is observed as the proportion of the experiments that result in the outcome.

Subjective Interpretation of Probability:

In the subjective interpretation, the probability of an outcome is not thought of as being a property of the outcome.

It is considered as a statement about the beliefs of the person who is quoting the probability.

The probability is about the chances of occurrence of the outcome.

Subjective Interpretation of Probability:

So probability becomes a subjective or personal concept and has no meaning outside of expressing one's degree of belief.

This interpretation of probability is often favored by decision makers.

Interpretation of Probability:

Irrespective of frequency or subjective interpretation of probability, practically there is a consensus that the mathematics of probability are the same in either case.

For example, if we say that the probability of raining tomorrow is 0.8, then we feel that

- 80% chances are there for the rain tomorrow and the expected weather will be cloudy.**
- 20% chances are there for the not raining tomorrow and the expected weather will not be cloudy.**

Essentials of Data Science With R Software - 1

Probability and Statistical Inference

Probability Theory

:::

Lecture 9

Sample Space and Events

Shalabh

Department of Mathematics and Statistics

Indian Institute of Technology Kanpur

Experiment:

Any activity for which the outcome is uncertain can be thought of as an “experiment.”

The uncertainty concerns in the sense that the outcome of the experiment is not known until the experiment is completed.

The outcome is not predictable with certainty in advance.

For example:

- **drawing a card from a deck to observe which card is drawn ,**
- **tossing a coin to observe what turns up – Head or Tail.**

Sample Space and Events:

The outcome of the experiment is not known in advance, but we assume that all the possible outcomes of the experiment are known.

This set of all possible outcomes of an experiment is known as the sample space of the experiment and is denoted by Ω .

Any subset E of the sample space is known as an event.

An event is a set of possible outcomes of the experiment.

If the outcome of the experiment is contained in E , then we say that E has occurred.

Sample Space and Events:

For example:

1. If the outcome of an experiment consists in the determination of the gender of a newly born child, then

$$\Omega = \{M, F\}$$

where M and F indicates Male and Female child, respectively.

If $E = \{M\}$, then E is the event that the child is a male (boy).

Similarly, if $E = \{F\}$, then E is the event that the child is a female (girl).

Sample Space and Events:

For example:

2. If three students have to get three positions in a game – First (1), Second (2) and Third (3). Then the experiment consists of the participating in the game with positions 1, 2, and 3, then

$$\Omega = \{\text{all orderings of } (1, 2, 3)\}$$

The outcome (2, 3, 1) means student one gets position 2, student two gets position 3 and student one gets position 1 and so on.

If $E = \{3, 2, 1\}$ then student one gets position 3, student two gets position 2 and student one gets position 1.

Sample Space and Events:

For example:

3. An experiment is conducted to know the dosage of a medicine. The dosage is increased continuously until a patient reacts positively.

One possible sample space for this experiment is to let Ω consist of all the positive numbers, so

$$\Omega = (0, \infty)$$

where the outcome would be x if the patient starts getting the dosage and reacts when the value of dosage reaches x .

No reaction to any smaller dosage than x .

Sample Space and Events: Symbols and Notations

Outcome of a random experiment is called a simple event (or elementary event) and denoted by ω .

Sample space : $\Omega = \{\omega_1, \omega_2, \dots, \omega_k\}$ is the set of all possible outcomes $\{\omega_1, \omega_2, \dots, \omega_k\}$.

Subsets of Ω are called events and are denoted by capital letters, in general, such as A, B, C .

Sample Space and Events: Complementary and Sure Events

Ω_A : Set of all simple events that are contained in the event A

The event \bar{A} refers to the non-occurring of A and is called a **composite or complementary event**.

Also Ω is an event.

Since it contains all possible outcomes, we say that Ω will always occur and is called a **sure event or certain event**.

Sample Space and Events: Impossible Event

On the other hand, if we consider the null set $\emptyset = \{\}$ as an event, then this event can never occur and is called an **impossible event**.

The sure event therefore is the set of all elementary events, and the impossible event is the set with no elementary events.

Sample Space and Events: Example - Rolling a die:

Rolling a die: If a die is rolled once, then the possible outcomes are the number of dots on the upper surface: 1, 2, . . . , 6.

Sample space is the set of simple events

$$\omega_1 = \text{"1"}, \quad \omega_2 = \text{"2"}, \quad \omega_3 = \text{"3"}, \quad \omega_4 = \text{"4"}, \quad \omega_5 = \text{"5"}, \quad \omega_6 = \text{"6"}.$$

$$\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5, \omega_6\}.$$

Event A: “An even number of dots on the upper surface of the die”.

There are three possibilities that this event occurs: $\omega_2, \omega_4, \omega_6$.

Complementary event of A: \bar{A} : If an odd number shows up.

There are three possibilities that this event occurs: $\omega_1, \omega_3, \omega_5$.

Sample Space and Events: Example - Rolling a die:

Elementary event is an event defined to observe only one particular number, say $\omega_1 = "1"$, then it is an elementary event.

Sure event is the event that “a number which is greater than or equal to 1” because any number between 1 and 6 is greater than or equal to 1.

Impossible event is the event that “the number is 7”.

Sample Space and Events: Example - Rolling two dice

Rolling two dice : Suppose we throw two dice simultaneously.

Event is defined as the “number of dots observed on the upper surface of both the dice”.

Then, there are 36 simple events defined as

(number of dots on first die, number of dots on second die),

i.e. $\omega_1 = (1, 1), \omega_2 = (1, 2), \dots, \omega_{36} = (6, 6)$.

Sample Space and Events: Example - Rolling two dice

Therefore Ω is

$$\begin{aligned}\Omega = & \{(1, 1), (1, 2), (1, 3), (1, 4), (1, 5), (1, 6) \\ & (2, 1), (2, 2), (2, 3), (2, 4), (2, 5), (2, 6) \\ & (3, 1), (3, 2), (3, 3), (3, 4), (3, 5), (3, 6) \\ & (4, 1), (4, 2), (4, 3), (4, 4), (4, 5), (4, 6) \\ & (5, 1), (5, 2), (5, 3), (5, 4), (5, 5), (5, 6) \\ & (6, 1), (6, 2), (6, 3), (6, 4), (6, 5), (6, 6)\}.\end{aligned}$$

Sample Space and Events: Example - Rolling two dice

One can define different events and their corresponding sample spaces.

For example,

- if an event A is defined as “upper faces of both the dice contain the same number of dots”, then the sample space is

$$\Omega_A = \{(1, 1), (2, 2), (3, 3), (4, 4), (5, 5), (6, 6)\}.$$

- If another event B is defined as “the sum of numbers on the upper faces is 6”, then the sample space is

$$\Omega_B = \{(1, 5), (2, 4), (3, 3), (4, 2), (5, 1)\}.$$

Sample Space and Events: Example - Rolling two dice

A sure event is “get either an even number or an odd number”

An impossible event would be “the sum of the two dice is greater than 13”.

Essentials of Data Science With R Software - 1

Probability and Statistical Inference

Probability Theory

:::

Lecture 10

Set Theory and Events Using Venn Diagrams

Shalabh

Department of Mathematics and Statistics

Indian Institute of Technology Kanpur

Union of Events:

Suppose A and B are any two events of a sample space Ω .

Define a new event $A \cup B$

$A \cup B$ is called the union of the events A and B .

$A \cup B$ consist of all outcomes that are

- either in A
- or in B
- or in both A and B .

Union of Events:

For example:

If the outcome of an experiment consists in the determination of the gender of a newly born child, then $\Omega = \{M, F\}$ where M and F indicates Male and Female child, respectively.

If $A = \{M\}$, then A is the event that the child is a male (boy).

Similarly, if $B = \{F\}$, then B is the event that the child is a female (girl).

Then $A \cup B = \{M, F\}$, i.e., $A \cup B$ is the whole sample space Ω .

$\Omega = A \cup B$ is called as sure event

Intersection of Events:

Suppose A and B are any two events of a sample space Ω .

Define a new event $A \cap B$

$A \cap B$ is called the intersection of the events A and B .

$A \cap B$ consist of all outcomes that are in both A and B .

Event $A \cap B$ will occur if both A and B occur.

Events with Venn Diagram using Set Theory:

It is possible to view events as sets of simple events.

This helps to determine how different events relate to each other.

A popular technique to visualize this approach is to use Venn diagrams.

Events with Venn Diagram using Set Theory:

In Venn diagrams, two or more sets are visualized by circles.

Overlapping circles

Separated circles

Events with Venn Diagram using Set Theory:

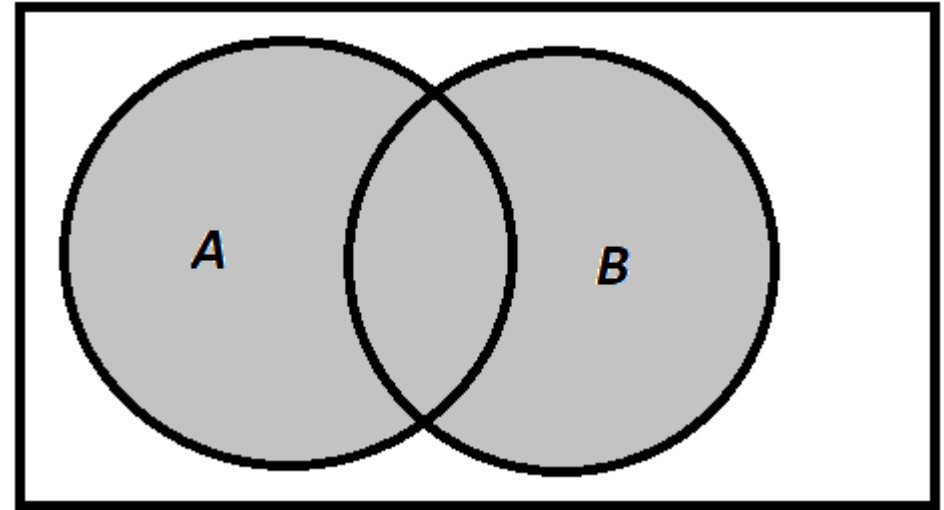
We use the following notations:

$A \cup B$: The union of events

$A \cup B$ is the set of all simple events

A and B which occur when a simple

Events A or B occurs (grey shaded area in figure).



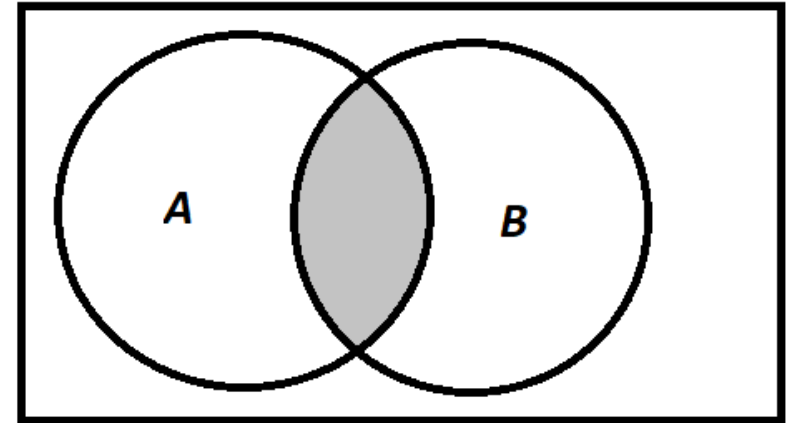
Please note that we use the word “or” from a statistical perspective:

“ A or B ” means that either a simple event from A occurs, or a simple event from B occurs, or a simple event which is part of both A and B occurs.

Events with Venn Diagram using Set Theory:

We use the following notations:

$A \cap B$: The intersection of events $A \cap B$ is the set of all simple events of A and B which occurs when the simple events of A and B occur (grey shaded area in figure).



Please note that we use the word “and” from a statistical perspective: “ A and B ” means that both simple events from A and from B occur.

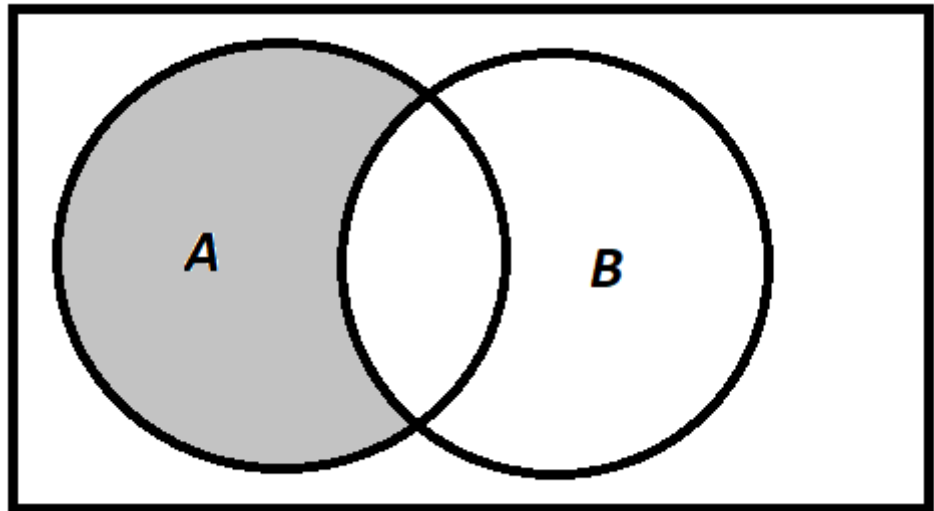
$A \cap B$ is also represented as AB .

Events with Venn Diagram using Set Theory:

We use the following notations:

$A - B$: The event $A - B$ contains all simple events of A , which are not contained in B .

(grey shaded area in figure).



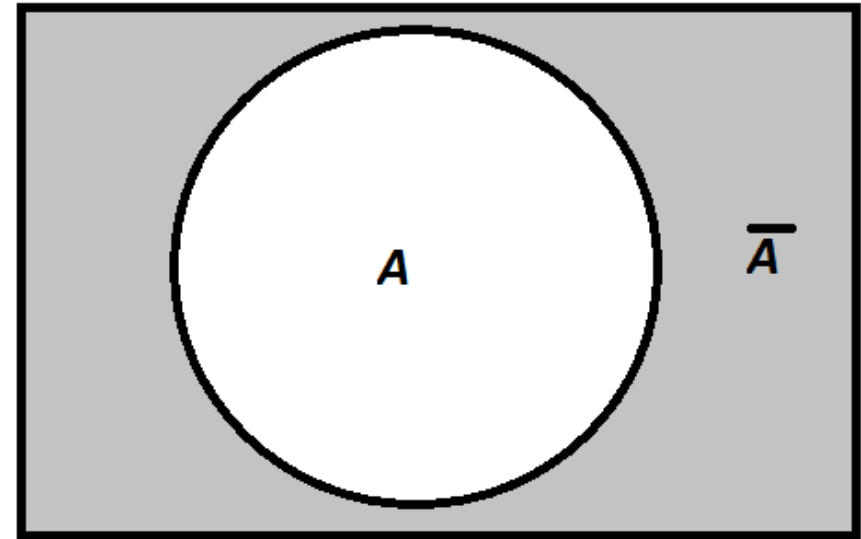
The event “ A but not B ” or “ A minus B ” occurs, if A occurs but B does not occur.

Also $A - B = A \cup \bar{B}$

Events with Venn Diagram using Set Theory:

We use the following notations:

\bar{A} : The event \bar{A} contains all simple events of Ω , which are not contained in A .

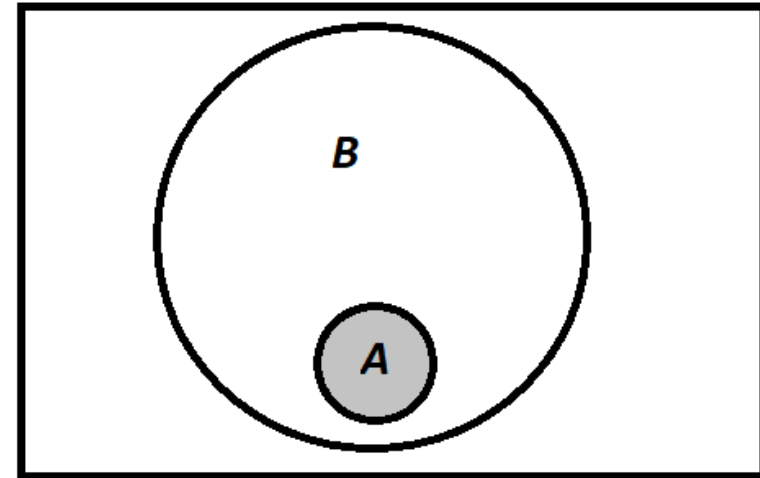


The complementary event of A (which is “Not- A ” or “ \bar{A} ” occurs whenever A does not occur (grey shaded area in figure)

Events with Venn Diagram using Set Theory:

We use the following notations:

$A \subseteq B$: A is a subset of B . This means
That all simple events of A are also
part of the sample space of B .



Events with Venn Diagram : Example - Rolling a die

Rolling a die: If a die is rolled once, then the possible outcomes are the number of dots on the upper surface: 1, 2, . . . , 6.

Sample space is the set of simple events

$$\omega_1 = \text{"1"}, \quad \omega_2 = \text{"2"}, \quad \omega_3 = \text{"3"}, \quad \omega_4 = \text{"4"}, \quad \omega_5 = \text{"5"}, \quad \omega_6 = \text{"6"}.$$

$$\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5, \omega_6\}.$$

- If $A = \{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5\}$ and B is the set of all odd numbers, then $B = \{\omega_1, \omega_3, \omega_5\}$ and thus $B \subseteq A$.

Events with Venn Diagram using Set Theory: Example - Rolling a die:

- If $A = \{\omega_2, \omega_4, \omega_6\}$ is the set of even numbers and $B = \{\omega_3, \omega_6\}$ is the set of all numbers which are divisible by 3, then $A \cup B = \{\omega_2, \omega_3, \omega_4, \omega_6\}$ is the collection of simple events for which the number is either even or divisible by 3 or both.

Events with Venn Diagram using Set Theory: Example - Rolling a die:

- If $A = \{\omega_1, \omega_3, \omega_5\}$ is the set of odd numbers and
 $B = \{\omega_3, \omega_6\}$ is the set of the numbers which are divisible by 3,
then $A \cap B = \{\omega_3\}$ is the set of simple events in which the numbers are odd and divisible by 3.
- If $A = \{\omega_1, \omega_3, \omega_5\}$ is the set of odd numbers and
 $B = \{\omega_3, \omega_6\}$ is the set of the numbers which are divisible by 3,
then $A - B = \{\omega_1, \omega_5\}$ is the set of simple events in which the numbers are odd but not divisible by 3.

Events with Venn Diagram using Set Theory: Example - Rolling a die:

- If $A = \{\omega_2, \omega_4, \omega_6\}$ is the set of even numbers, then
 $\bar{A} = \{\omega_1, \omega_3, \omega_5\}$ is the set of odd numbers.

Disjoint Events with Set Theory

Two events A and B are **disjoint** if $A \cap B = \emptyset$ holds,
i.e. if both events cannot occur simultaneously.

Example:

The events A and \bar{A} are disjoint events.

Mutually Disjoint Events with Set Theory

The events A_1, A_2, \dots, A_m are said to be mutually or pairwise disjoint, if $A_i \cap A_j = \emptyset$ whenever $i \neq j = 1, 2, \dots, m$.

Example: Rolling a die: If a die is rolled once, then the possible outcomes are the number of dots on the upper surface: $1, 2, \dots, 6$.

$$\omega_1 = \text{"1"}, \quad \omega_2 = \text{"2"}, \quad \omega_3 = \text{"3"}, \quad \omega_4 = \text{"4"}, \quad \omega_5 = \text{"5"}, \quad \omega_6 = \text{"6"}.$$

$$\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5, \omega_6\}.$$

If $A = \{\omega_1, \omega_3, \omega_5\}$ and $B = \{\omega_2, \omega_4, \omega_6\}$ are the sets of odd and even numbers, respectively, then the events A and B are disjoint.

Unions of More than Two Events:

We can also define unions of more than two events.

Union of the events A_1, A_2, \dots, A_m , denoted by $A_1 \cup A_2 \cup \dots \cup A_m$ is defined to be the event consisting of all outcomes that are in A_i for at least one $i = 1, 2, \dots, m$.

In other words, the union of the A_i occurs when at least one of the events A_i occurs.

Intersections of More than Two Events:

We can also define intersections of more than two events.

Intersection of the events A_1, A_2, \dots, A_m ,

denoted by $A_1 \cap A_2 \cap \dots \cap A_m$ is defined to be the event consisting of those outcomes that are in all of the events $A_i, i = 1, 2, \dots, m$.

In other words, the intersection occurs when all of the events A_i occur.

Essentials of Data Science With R Software - 1

Probability and Statistical Inference

Probability Theory

:::

Lecture 12

Probability and Relative Frequency- An Example

Shalabh

Department of Mathematics and Statistics

Indian Institute of Technology Kanpur

Relative Frequency and Probability of an Event: Example- Dice Roll

Suppose a fair dice is rolled and its outcome as the number of points on the upper face is recorded as 1, 2, 3, 4, 5, 6.

Sample space $(\Omega) = \{1, 2, 3, 4, 5, 6\}$

Suppose we repeat the experiment 100 times and the outcomes are recorded and the relative frequencies are obtained.

Relative Frequency and Probability of an Event: Example- Dice Roll

Suppose we repeat the experiment 100 times and the outcomes are recorded and the relative frequencies are obtained as follows:

Total number of 1's = 15 $f(1) = 15/100$

Total number of 2's = 10 $f(2) = 10/100$

Total number of 3's = 25 $f(3) = 25/100$

Total number of 4's = 14 $f(4) = 14/100$

Total number of 5's = 16 $f(5) = 16/100$

Total number of 6's = 20 $f(6) = 20/100$

Relative Frequency and Probability of an Event: Example-Dice Roll

Meaning of a fair dice: Probabilities of observing 1, 2, 3, 4, 5, 6 are equal, i.e., $1/6$.

When the fair dice is rolled a large number of times and n tends to infinity, then all $f(A_i)$, $i = 1, 2, \dots, 6$ will have a limiting value $1/6$ which is the probability of getting 1, 2, 3, 4, 5, or 6.

Relative Frequency and Probability of an Event: Example- Dice Roll

This can be simulated in R by the `sample` command by drawing the observations among 1, 2, 3, 4, 5, 6 by simple random sampling with replacement and then finding the relative frequencies using the `table` and `length` commands.

Suppose we want repeat the experiment 100 times. This means drawing 100 values and finding the relative frequencies of 1, 2, 3, 4, 5, and 6.

Relative Frequency and Probability of an Event: Example-Dice Roll

The command

```
dice100 = sample(c(1,2,3,4,5,6), size=100,  
replace = T)
```

generates 100 values and stores it in a data vector `dice100`.

Then the following command computes the relative frequencies of the data stored in `dice100`:

```
table(dice100)/length(dice100)
```

So we repeat by increasing the number of repetitions $n = 10, 100, 1000, 10000, \dots$

Relative Frequency and Probability of an Event: Example- Dice Roll

100 repetitions

```
> dice100 = sample(c(1,2,3,4,5,6), size=100,  
replace = T)
```

```
> table(dice100)/length(dice100)
```

1	2	3	4	5	6
0.10	0.17	0.15	0.23	0.14	0.21

```
> dice100 = sample(c(1,2,3,4,5,6), size=100,  
replace = T)
```

```
> table(dice100)/length(dice100)
```

1	2	3	4	5	6
0.13	0.18	0.22	0.14	0.16	0.17

Relative Frequency and Probability of an Event: Example-Dice Roll

1000 repetitions

```
> dice1000 = sample(c(1,2,3,4,5,6), size=1000,  
replace = T)
```

```
> table(dice1000)/length(dice1000)
```

1	2	3	4	5	6
0.147	0.169	0.180	0.181	0.155	0.168

```
> table(dice1000)/length(dice1000)
```

1	2	3	4	5	6
0.175	0.180	0.174	0.163	0.171	0.137

Relative Frequency and Probability of an Event: Example-Dice Roll

1000 repetitions

```
> dice10000 = sample(c(1,2,3,4,5,6), size=10000,  
replace = T)
```

```
> table(dice10000)/length(dice10000)
```

1	2	3	4	5	6
0.1626	0.1680	0.1657	0.1683	0.1718	0.1636

```
> table(dice10000)/length(dice10000)
```

1	2	3	4	5	6
0.1626	0.1680	0.1657	0.1683	0.1718	0.1636

Relative Frequency and Probability of an Event: Example- Dice Roll

```
R Console

dice100
  1    2    3    4    5    6
0.10 0.17 0.15 0.23 0.14 0.21
>
> dice100 = sample(c(1,2,3,4,5,6), size=100, replace = T) # 100 repetitions
> table(dice100)/length(dice100) # Relative frequencies
dice100
  1    2    3    4    5    6
0.13 0.18 0.22 0.14 0.16 0.17
>
> dice1000 = sample(c(1,2,3,4,5,6), size=1000, replace = T) # 1000 repetitions
> table(dice1000)/length(dice1000) # Relative frequencies
dice1000
  1    2    3    4    5    6
0.147 0.169 0.180 0.181 0.155 0.168
> table(dice1000)/length(dice1000) # Relative frequencies
dice1000
  1    2    3    4    5    6
0.175 0.180 0.174 0.163 0.171 0.137
>
> dice10000 = sample(c(1,2,3,4,5,6), size=10000, replace = T) # 10000 repetitions
> table(dice10000)/length(dice10000) # Relative frequencies
dice10000
  1    2    3    4    5    6
0.1626 0.1680 0.1657 0.1683 0.1718 0.1636
> table(dice10000)/length(dice10000) # Relative frequencies
dice10000
  1    2    3    4    5    6
0.1626 0.1680 0.1657 0.1683 0.1718 0.1636
>
```

Essentials of Data Science With R Software - 1

Probability and Statistical Inference

Probability Theory

:::

Lecture 13

Axiomatic Definition of Probability

Shalabh

Department of Mathematics and Statistics

Indian Institute of Technology Kanpur

Relative Frequency and Probability of an Event:

If we assume that

- the experiment is repeated a large number of times
(mathematically, this would mean that n tends to infinity) and
- the experimental conditions remain the same (at least approximately) over all the repetitions,

then the relative frequency $f(A)$ converges to a limiting value for A .

This limiting value is interpreted as the probability of A and denoted by

$$P(A) = \lim_{n \rightarrow \infty} \frac{n(A)}{n}$$

where $n(A)$ denotes the number of times an event A occurs out of n times.

Relative Frequency and Probability of an Event:

Although this definition is certainly intuitively pleasing and but it possesses a serious drawback:

How do we know that $n(A)/n$ will converge to some constant limiting value that will be the same for each possible sequence of repetitions of the experiment?

Relative Frequency and Probability of an Event:

For example, a coin is continuously tossed repeatedly.

1. How do we know that the proportion of heads obtained in the first n tosses will converge to some value as n gets large?
2. Even if it converges to some value, how do we know that, if the experiment is repeatedly performed a second time, we will again obtain the same limiting proportion of heads?

Relative Frequency and Probability of an Event:

This issue is answered by stating the convergence of $n(A)/n$ to a constant limiting value as an assumption, or an axiom, of the system.

However, to assume that $n(A)/n$ will necessarily converge to some constant value is a complex assumption.

We hope that such a constant limiting frequency exists, it is difficult to believe a priori that this will happen.

Relative Frequency and Probability of an Event:

In fact, it would be better to assume a set of simpler axioms about probability and then attempt to prove that such a constant limiting frequency does in some sense exist.

This approach is the modern axiomatic approach to probability theory.

Relative Frequency and Probability of an Event:

We assume that for each event A in the sample space Ω there exists a value $P(A)$, referred to as the probability of A .

We then assume that the probabilities satisfy a certain set of axioms which will be more agreeable with our intuitive notion of probability.

Axiomatic Definition of Probability:

From a purely mathematical viewpoint, we suppose that for each event A of an experiment having a sample space Ω there is a number, denoted by $P(A)$ which satisfies the following three axioms:

Axiom 1: Every random event A has a probability in the (closed) interval $[0, 1]$, i.e., $0 \leq P(A) \leq 1$

Axiom 2: The sure event has probability 1, i.e., $P(\Omega) = 1$

Axiom 3: For any sequence of disjoint or mutually exclusive events $A_1, A_2, \dots, A_n, \dots$, (that is, events for which $A_i \cap A_j = \emptyset$ when $i \neq j$),
$$P(A_1 \cap A_2 \cap \dots \cap A_n \cap \dots) = P(A_1) + P(A_2) + \dots + P(A_n) + \dots, n = 1, 2, \dots, \infty$$

We call $P(A)$ the probability of the event A .

Axiomatic Definition of Probability:

Axiom 1 states that the probability that the outcome of the experiment is contained in A is some number between 0 and 1.

Axiom 2 states that, with probability 1, the outcome will be a member of the sample space Ω .

Axiom 3 states that for any set of mutually exclusive events the probability that at least one of these events occurs is equal to the sum of their respective probabilities.

Axiom 3 is called the **theorem of additivity of disjoint events.**

Axiomatic Definition of Probability:

It is to be noted that if we interpret $P(A)$ as the relative frequency of the event A when a large number of repetitions of the experiment are performed, then $P(A)$ would indeed satisfy the above axioms.

For instance,

- the proportion (or frequency) of time that the outcome is in A is clearly between 0 and 1, and**
- the proportion of time that it is in Ω is 1 (since all outcomes are in Ω).**
- Also, if A and B have no outcomes in common, then the proportion of time that the outcome is in either A or B is the sum of their respective frequencies.**

Axiomatic Definition of Probability:

Example: Suppose a pair of dice is rolled and sum of the points on upper faces is obtained.

Suppose event A : sum is 4, 6, or 12 and
event B is that the sum is 7 or 9.

Then if outcome A occurs 10% time and outcome B occurs 20% time,
then 30% of the time the outcome will be either 4, 6, 12, 7, or 9.

Essentials of Data Science With R Software - 1

Probability and Statistical Inference

Probability Theory

:::

Lecture 14

Some Rules of Probability

Shalabh

Department of Mathematics and Statistics

Indian Institute of Technology Kanpur

Some Rules of Probability:

- The probability of occurrence of an impossible event ϕ is zero:

$$P(\phi) = 1 - P(\Omega) = 0.$$

- The probability of occurrence of a sure event is one:

$$P(\Omega) = 1.$$

- The probability of the complementary event of A , (i.e. \bar{A}) is

$$P(\bar{A}) = 1 - P(A).$$

Some Rules of Probability:

- The *odds* of an event A is defined by

$$\frac{P(A)}{P(\bar{A})} = \frac{P(A)}{1 - P(A)}$$

Thus the odds of an event A tells how much more likely it is that A occurs than that it does not occur.

Example: if $P(A) = 3/4$, then $\frac{P(A)}{1 - P(A)} = 3$, so the odds are 3.

Consequently, it is 3 times as likely that A occurs as it is that it does not.

Some Rules of Probability:

Example: Suppose a box of 30 ice creams contains ice creams of 6 different flavours with 5 ice creams of each flavour.

Suppose an event A is defined as $A = \{\text{"vanilla flavour"}\}$.

Probability of finding a vanilla flavour ice cream (without looking into the box) = $P(\text{"vanilla flavour"}) = 5/30$.

Then, the probability of the complementary event \bar{A} , i.e. the probability of not finding a vanilla flavour ice cream is $P(\text{"no vanilla flavour"}) = 1 - P(\text{"vanilla flavour"}) = 25/30$.

Some Rules of Probability: Additive Theorem of Probability

Let A_1 and A_2 be not necessarily disjoint events.

The probability of occurrence of A_1 or A_2 is

$$P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2).$$

The meaning of “or” is in the statistical sense: either A_1 is occurring, A_2 is occurring, or both of them.

Some Rules of Probability: Additive Theorem of Probability

Example: A total of 28% people like sweet snacks, 7% like salty snacks, and 5% like both sweet and salty snacks. The percentage of people like neither sweet nor salty snacks is obtained as follows:

Let A_1 be the event that a randomly chosen person likes sweet snacks and A_2 be the event that a randomly chosen person likes salty snacks.

Then, the probability this person likes either sweet or salty snacks is

$$P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2) = 0.07 + 0.28 - 0.05 = 0.30$$

Thus 70% of people does not like either sweet or salty snacks.

Sample Spaces having Equally Likely Outcomes

For a large number of experiments, it is natural to assume that each point in the sample space is equally likely to occur.

For many experiments whose sample space Ω is a finite set, say

$\Omega = \{1, 2, \dots, N\}$, it is often natural to assume that

$$P(\{1\}) = P(\{2\}) = \dots = P(\{N\}) = p \text{ (say)}$$

$$P(\Omega) = P(\{1\}) + \dots + P(\{N\})$$

$$\text{or } 1 = Np$$

$$\text{or } P(\{i\}) = p = \frac{1}{N}$$

Sample Spaces having Equally Likely Outcomes

For any event A ,

$$P(A) = \frac{\text{Number of points in } A}{N}$$

In other words, if we assume that each outcome of an experiment is equally likely to occur, then

the probability of any event A = the proportion of points in the
sample space that are contained in A .

Thus, to compute probabilities it is necessary to know the number of different ways to count given events.

Essentials of Data Science With R Software - 1

Probability and Statistical Inference

Probability Theory

:::

Lecture 15

Basic Principle of Counting-Ordered Set, Unordered Set and Permutations

Shalabh

Department of Mathematics and Statistics

Indian Institute of Technology Kanpur

Sample Spaces having Equally Likely Outcomes

For any event A ,

$$P(A) = \frac{\text{Number of points in } A}{N}$$

In other words, if we assume that each outcome of an experiment is equally likely to occur, then

the probability of any event A = the proportion of points in the
sample space that are contained in A .

Thus, to compute probabilities it is necessary to know the number of different ways to count given events.

Basic Principle of Counting:

Suppose that two experiments are to be performed.

Suppose experiment 1 can result in any one of m possible outcomes and if, for each outcome of experiment 1, there are n possible outcomes of experiment 2,

then together there are mn possible outcomes of the two experiments.

$(1, 1), (1, 2), \dots, (1, n)$

$(2, 1), (2, 2), \dots, (2, n)$

...

$(m, 1), (m, 2), \dots, (m, n)$

Ordered and Unordered Sets:

Suppose three balls of different colours, black, grey, and white, are drawn.

Now there are two options:

1. The first option is to take into account the order in which the balls are drawn.

In such a situation, two possible sets of balls such as (black, grey, and white) and (white, black, and grey) constitute two different sets.

Such a set is called an *ordered set*.

Ordered and Unordered Sets:

2. In the second option, we do not take into account the order in which the balls are drawn.

In such a situation, the two possible sets of balls such as

(black, grey, and white) and (white, black, and grey)

are the same sets and constitute an *unordered set* of balls.

A group of elements is said to be *ordered* if the order in which these elements are drawn is of relevance.

Otherwise, it is called *unordered*.

Ordered and Unordered Sets: Examples

- **Ordered samples:**

- The first three places in a 100m race are determined by the order in which the athletes arrive at the finishing line.

If 8 athletes are competing with each other, the number of possible results for the first three places is of interest.

- In a lottery with two prizes, the first drawn lottery ticket gets the first prize and the second lottery ticket gets the second prize.

Ordered and Unordered Sets: Examples

- **Unordered samples:**

- The selected members for a football team. The order in which the selected names are announced is irrelevant.

- Fishing 20 fish from a lake.

- A bunch of 10 flowers made from 21 flowers of 4 different colours

Factorial function:

The factorial function $n!$ is defined as

$$n! = 1 \times 2 \times 3 \times \cdots \times n \text{ for } n > 0 \quad \text{and} \quad 0! = 1.$$

Thus $1! = 1$

$$2! = 1 \times 2 = 2,$$

$$3! = 1 \times 2 \times 3 = 6 .$$

This can be calculated in R as follows:

```
factorial(n)
```

Factorial function in R:

Factorial function can be calculated in R as follows:

`factorial(n)`

Example:

```
> factorial(0)
```

```
[1] 1
```

```
> factorial(1)
```

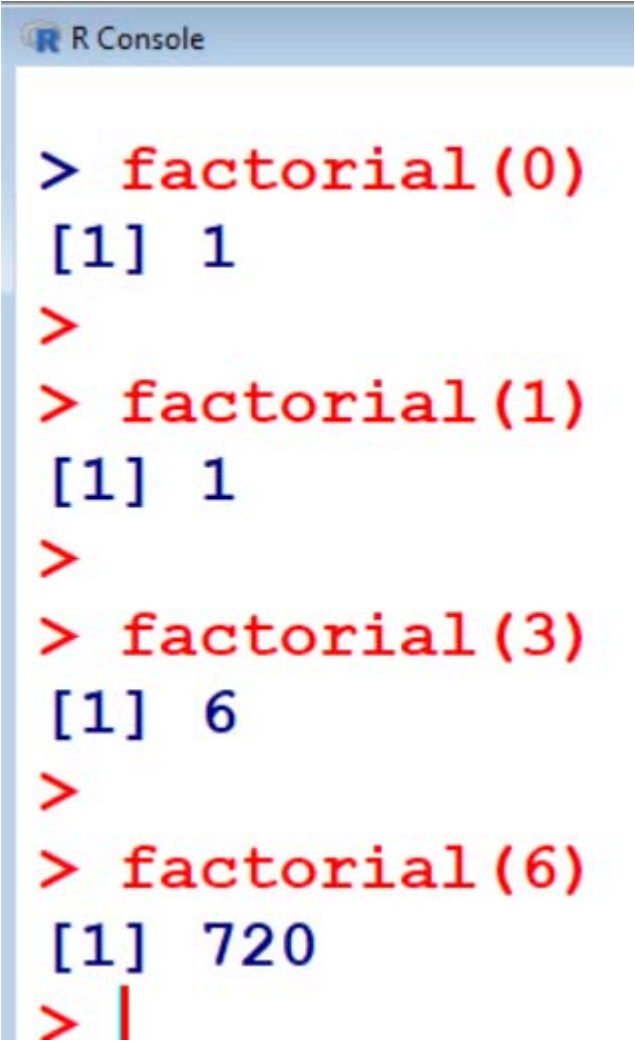
```
[1] 1
```

```
> factorial(3)
```

```
[1] 6
```

```
> factorial(6)
```

```
[1] 720
```

A screenshot of the R Console window. The title bar says "R Console". The console shows several lines of code and output. The code is in red, and the output is in blue. The code includes: > factorial(0), > factorial(1), > factorial(3), and > factorial(6). The output shows: [1] 1, [1] 1, [1] 6, and [1] 720. There is a vertical cursor at the end of the last line of code.

```
> factorial(0)
[1] 1
>
> factorial(1)
[1] 1
>
> factorial(3)
[1] 6
>
> factorial(6)
[1] 720
> |
```

Permutation:

Consider a set of n elements.

Each ordered composition of these n elements is called a permutation.

We distinguish between two cases:

- If all the elements are distinguishable, then we speak of permutation without replacement.
- If some or all of the elements are not distinguishable, then we speak of permutation with replacement.

Note: the meaning of “replacement” is just a convention and does not directly refer to the drawings.

Permutations Without Replacement :

Consider a set of n elements.

If all the n elements are distinguishable, then there are $n!$ different compositions of these elements.

Example: There are 3 students who will get three ranks – First (F), Second (S) and Third (T).

There are $3! = 6$ possible ways in which they can be ranked.

(F, S, T), (F, T, S), (S, T, F), (S, F, T), (T, F, S), (T, S, F)

Permutations Without Replacement :

Example: A person has 10 books that he is going to put on his bookshelf. Of these,

- 4 are mathematics books,
- 3 are chemistry books,
- 2 are history books, and
- 1 is a language book.

He wants to arrange his books so that all the books dealing with the same subject are together on the shelf.

We want to know the total number of possible different arrangements.

Permutations Without Replacement :

Solution:

There are $4! \ 3! \ 2! \ 1!$ arrangements such that the mathematics books are first in line, then the chemistry books, then the history books, and then the language book.

Similarly, for each possible ordering of the subjects, there are $4! \ 3! \ 2! \ 1!$ possible arrangements.

Hence, as there are $4!$ possible orderings of the subjects, the desired answer is

$$\mathbf{4! \ 4! \ 3! \ 2! \ 1! = 6,912.}$$

Permutations With Replacement :

Consider a set of n elements.

Assume that not all n elements are distinguishable.

The elements are divided into groups, and these groups are distinguishable.

Suppose, there are s groups of sizes n_1, n_2, \dots, n_s .

The total number of different ways to arrange the n elements in s groups is

$$\frac{n!}{n_1!n_2!\cdots n_s!}.$$

Permutations With Replacement :

Example: There were 10 students and there are 3 types of chocolates- C1, C2 and C3. The total number of ways in which two C1, three C2 and five C3 chocolates can be given to the 10 students is obtained as follows:

$$n_1 = 2, n_2 = 3, n_3 = 5, n = 10$$

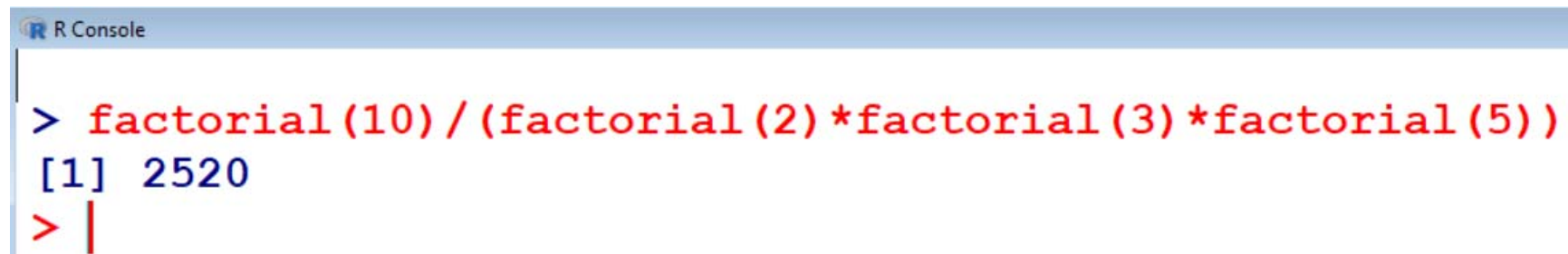
The total number of different ways to arrange the $n = 10$ elements in 3 groups is

$$\frac{10!}{2!3!5!}$$

Factorial function:

This can be calculated in R as follows:

```
factorial(10)/(factorial(2)*factorial(3)*factorial(5))  
[1] 2520
```

A screenshot of an R console window. The title bar at the top says "R Console". The console shows a command being entered: "> factorial(10)/(factorial(2)*factorial(3)*factorial(5))". The output is displayed below the command: "[1] 2520". A red prompt character ">" and a vertical cursor line are visible on the next line.

```
R Console  
> factorial(10)/(factorial(2)*factorial(3)*factorial(5))  
[1] 2520  
> |
```

Essentials of Data Science With R Software - 1

Probability and Statistical Inference

Probability Theory

:::

Lecture 16

Basic Principle of Counting- Combinations

Shalabh

Department of Mathematics and Statistics

Indian Institute of Technology Kanpur

Combinations:

The Binomial coefficient for any integers m and n with $n \geq m \geq 0$ is denoted and defined as

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

It is read as “ n choose m ” and can be calculated in R using the following command:

`choose(n,m)`

Combinations:

We now answer the question of how many different possibilities exist to draw m out of n elements, i.e. m out of n balls from an urn.

It is necessary to distinguish between the following four cases:

1. **Combinations without replacement and without consideration of the order of the elements.**
2. **Combinations without replacement and with consideration of the order of the elements.**
3. **Combinations with replacement and without consideration of the order of the elements.**
4. **Combinations with replacement and with consideration of the order of the elements.**

1. Combinations without replacement and without consideration of the order of the elements:

When there is no replacement and the order of the elements is also not relevant, then the total number of distinguishable combinations in drawing m out of n elements is

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

This result can be obtained in *R* by using the command

`choose(n,m)`

1. Combinations without replacement and without consideration of the order of the elements:

Example: Suppose a company elects a new board of directors. The board consists of 6 members and 10 people are eligible to be elected. How many combinations for the board of directors exist?

Since a person cannot be elected twice, we have a situation where there is no replacement. The order is also of no importance: either one is elected or not.

$$\binom{10}{6} = \frac{10!}{6! (10 - 6)!} = 210$$

possible combinations.

1. Combinations without replacement and without consideration of the order of the elements:

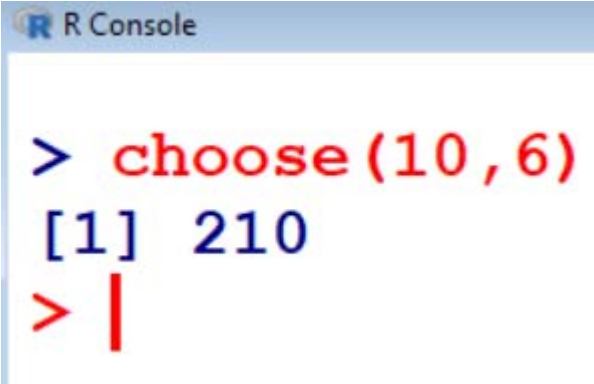
Example: (Contd.)

This result can be obtained in *R* by using the command

`choose(10,6)`

```
> choose(10,6)
```

```
[1] 210
```



```
R Console  
> choose(10,6)  
[1] 210  
> |
```

2. Combinations without replacement and with consideration of the order of the elements:

The total number of different combinations for the setting without replacement and with consideration of the order is

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

This can be calculated in *R* as follows:

```
factorial(n)/factorial(n-m)
```

or

```
factorial(m)*choose(n,m)
```

2. Combinations without replacement and with consideration of the order of the elements:

Example:

Consider a race with 10 students. A possible bet is to forecast the winner of the race, the second student of the race, and the third student of the race.

The total number of different combinations for the students in the first three places is $\frac{10!}{(10-3)!}$.

2. Combinations without replacement and with consideration of the order of the elements:

Example: (Contd.) This result can be explained intuitively:

- For the first place, there is a choice of 10 different students.
- For the second place, there is a choice of 9 different students (10 students minus the winner).
- For the third place, there is a choice of 8 different students (10 students minus the first and second students).
- The total number of combinations is $10 \times 9 \times 8$.

2. Combinations without replacement and with consideration of the order of the elements:

Example: (Contd.) This can be calculated in *R* as follows:

`10 * 9 * 8`

or

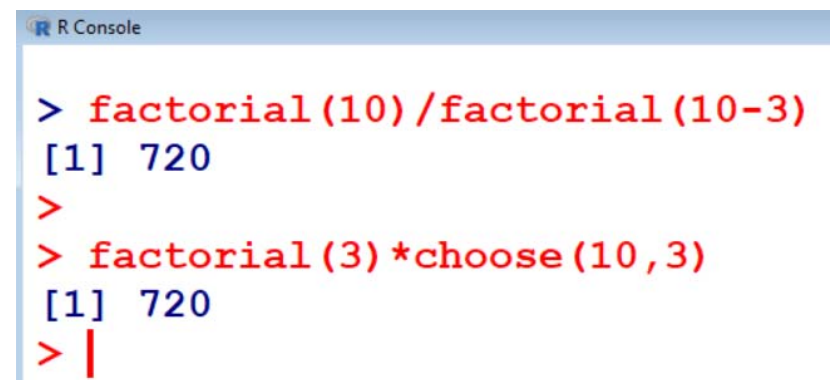
`factorial(10)/factorial(10-3)`

`[1] 720`

or

`factorial(3)*choose(10,3)`

`[1] 720`



```
R Console
> factorial(10)/factorial(10-3)
[1] 720
>
> factorial(3)*choose(10,3)
[1] 720
> |
```

3. Combinations with replacement and without consideration of the order of the elements:

The total number of different combinations with replacement and without consideration of the order is

$$\binom{n + m - 1}{m} = \frac{(n+m-1)!}{m!(n-1)!} = \binom{n + m - 1}{n - 1}.$$

This can be calculated in *R* as follows:

```
choose(n+m-1, m)
```


3. Combinations with replacement and without consideration of the order of the elements:

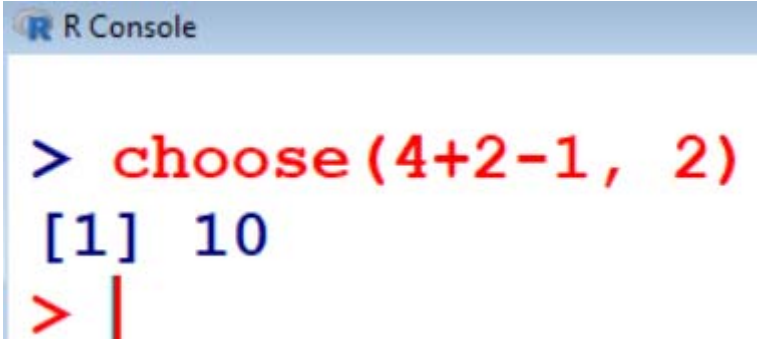
Example: A farmer has 2 fields and aspires to cultivate one out of 4 different organic products per field. Then, the total number of choices

he has is $\binom{4 + 2 - 1}{2} = 10$.

This can be calculated in *R* as follows:

```
choose(4+2-1, 2)
```

```
[1] 10
```

A screenshot of an R console window. The title bar says "R Console". The prompt ">" is followed by the command "choose(4+2-1, 2)" in red text. Below it, the output "[1] 10" is shown in blue text. The prompt ">" is followed by a vertical bar "|" in red text, indicating the cursor is at the end of the line.

```
> choose(4+2-1, 2)
[1] 10
> |
```

3. Combinations with replacement and without consideration of the order of the elements:

Example: (Contd.)

If 4 different organic products are denoted as a, b, c, and d, then the following combinations are possible:

(a, a) (a, b) (a, c) (a, d)

(b, b) (b, c) (b, d)

(c, c) (c, d)

(d, d)

Please note that, for example, (a, b) is identical to (b, a) because the order in which the products a and b are cultivated on the first or second field is not important in this example.

4. Combinations with replacement and with consideration of the order of the elements:

The total number of different combinations for the integers m and n with replacement and when the order is of relevance is

$$n^m .$$

This can be calculated in *R* as follows:

`n^m`

or

`n**m`

4. Combinations with replacement and with consideration of the order of the elements:

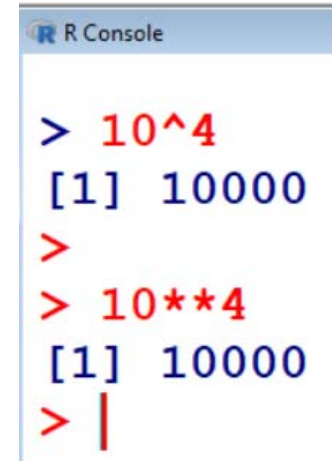
Example: Consider a credit card with a four-digit ATM personal identification number (PIN) code. The total number of possible combinations for the PIN is

$$n^m = 10^4 = 10000.$$

Note that every digit in the first, second, third, and fourth places ($m = 4$) can be chosen out of ten digits from 0 to 9 ($n = 10$).

This can be calculated in *R* as follows:

```
> 10^4  
[1] 10000  
> 10**4  
[1] 10000
```



```
R Console  
> 10^4  
[1] 10000  
>  
> 10**4  
[1] 10000  
> |
```

Example 1:

A committee of 5 is to be selected from a group of 6 men and 9 women. If the selection is made randomly, what is the probability that the committee consists of 3 men and 2 women?

Solution: Let us assume that randomly selected means that each of the $\binom{15}{5}$ possible combinations is equally likely to be selected. Hence the probability that committee consists of 3 men and 2 women

$$\frac{\binom{6}{3} \binom{9}{2}}{\binom{15}{5}} = \frac{240}{1001}$$

Example 2:

An urn contains n balls, of which one is special. If k of these balls are withdrawn one at a time, with each selection being equally likely to be any of the balls that remain at the time, what is the probability that the special ball is chosen?

Solution: Since all of the balls are treated in an identical manner, it follows that the set of k balls selected is equally likely to be any of the $\binom{n}{k}$ sets of k balls. Therefore,

$$P(\text{special ball is selected}) = \frac{\binom{1}{1} \binom{n-1}{k-1}}{\binom{n}{k}} = \frac{k}{n}$$

Example 3:

Following number of members of a club play tennis, squash and cricket:

	Tennis	Squash	Cricket
Number of players	36	28	18

Furthermore,

	Tennis and squash	Squash and cricket	Tennis and cricket	Tennis, squash and cricket
Number of players	22	9	12	4

How many members of this club play at least one of these sports?

Example 3:

Solution: Let

N denote the number of members of the club, and introduce probability by assuming that a member of the club is randomly selected.

If for any subset C of members of the club,

$P(C)$ denote the probability that the selected member is contained in C , then

$$P(C) = \frac{\text{number of members in } C}{N}$$

Example 3:

Now, with

T : Set of members that plays tennis,

S : Set that plays squash, and

C : Set that plays cricket, we have

$$P(T \cup S \cup C) = P(T) + P(S) + P(C) - P(TS) - P(TC) - P(SC) + P(TSC)$$

$$= \frac{36 + 28 + 18 - 22 - 12 - 9 + 4}{N} = \frac{43}{N}$$

Hence we can conclude that 43 members play at least one of the sports.

Essentials of Data Science With R Software - 1

Probability and Statistical Inference

Probability Theory

:::

Lecture 17

Conditional Probability

Shalabh

Department of Mathematics and Statistics

Indian Institute of Technology Kanpur

Conditional Probability:

Conditional probability is useful in calculating probabilities when some partial information concerning the result of the experiment is available, or in recalculating them in light of additional information.

In such situations, the desired probabilities are conditional ones.

Sometimes it is often the easiest way to compute the probability of an event is to first “condition” on the occurrence or non-occurrence of a secondary event.

Conditional Probability:

Consider the following example to understand the concept of conditional probability:

Suppose a blood test is developed to diagnose a particular infection. The blood test is conducted over 100 randomly selected persons.

The outcomes of the absolute and relative frequencies are presented in following Tables:

Conditional Probability:

Absolute frequencies of test results and infection status				
		Infection		Total (row)
		Present	Absent	
Test	Positive (+)	30	10	40
	Negative (-)	15	45	60
Total (Columns)		45	55	Total = 100

Conditional Probability:

Relative frequencies of test results and infection status				
		Infection		Total (row)
		Present (IP)	Absent (IA)	
Test	Positive (T+)	0.30	0.10	0.40
	Negative (T-)	0.15	0.45	0.60
Total (Columns)		0.45	0.55	Total = 1

Conditional Probability:

There are the following four possible outcomes:

1. The blood sample has an infection and the test diagnoses it, i.e. the test is correctly diagnosing the infection.

2. The blood sample does not has any infection and the test does not diagnose it, i.e. the test is correctly diagnosing that there is no infection.

Conditional Probability:

There are the following four possible outcomes:

3. The blood sample has an infection and the test does not diagnose it, i.e. the test is incorrect in stating that there is no infection.

4. The blood sample does not has any infection but the test diagnoses it, i.e. the test is incorrect in stating that there is an infection.

Conditional Probability:

If one already knows that the test is positive and wants to determine the probability that the infection is indeed present, then this can be achieved by the respective conditional probability $P(IP|T+)$ which is

$$P(IP|T+) = \frac{P(IP \cap T+)}{P(T+)} = \frac{0.3}{0.4} = 0.75$$

Note that $IP \cap T+$ denotes the “relative frequency of blood samples in which the disease is present and the test is positive” which is 0.3.

Conditional Probability:

Let $P(A) > 0$. Then the conditional probability of event B occurring, given that event A has already occurred, is

$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

The roles of A and B can be interchanged to define $P(A|B)$ as follows.

Let $P(B) > 0$. The conditional probability of A given B is

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Conditional Probability:

The definition of conditional probability is consistent with the interpretation of probability as being a long-run relative frequency, i.e., a large number n of repetitions of the experiment are performed.

Conditional Probability: Example 1

A coin is tossed twice. If we assume that all four points in the sample space $\Omega = \{(H, H), (H, T), (T, H), (T, T)\}$ are equally likely, what is the conditional probability that both tosses result in heads, given that the first toss results in head?

Solution:

If $A = \{(H, H)\}$ denotes the event that both tosses results in heads, and $B = \{(H, H), (H, T)\}$ the event that the first toss results in head, then the desired probability is

$$P(A/B) = \frac{P(A \cap B)}{P(B)} = \frac{P(\{(H, H)\})}{P(\{(H, H), (H, T)\})} = \frac{1/4}{2/4} = \frac{1}{2}$$

Conditional Probability: Example 2

An urn contains 10 white, 5 yellow, and 10 black marbles. A marble is chosen at random from the urn, and it is noted that it is not one of the black marbles. What is the probability that it is yellow?

Solution:

Let Y denote the event that the marble selected is yellow, and

\bar{B} denote the event that it is not black.

Now, then the desired probability is $P(Y|\bar{B})$

$$P(Y|\bar{B}) = \frac{P(Y \cap \bar{B})}{P(\bar{B})}$$

Conditional Probability: Example 2

However $Y \cap \bar{B} = Y$, since the marble will be both yellow and not black if and only if it is yellow.

Hence, assuming that each of the 25 marbles is equally likely to be chosen, we obtain that

$$P(Y|\bar{B}) = \frac{P(Y \cap \bar{B})}{P(\bar{B})} = \frac{5/25}{15/25} = \frac{1}{3}$$

Conditional Probability: Example 3

A box contains 5 defective, 10 partially defective (that fail after a couple of hours of use), and 25 acceptable (non-defective) transistors. A transistor is chosen at random from the box and put into use. If it does not immediately fail, what is the probability it is acceptable?

Conditional Probability: Example 3

Solution:

Since the transistor did not immediately fail, we know that it is not one of the 5 defectives and so the desired probability is:

$$P(\text{acceptable} | \text{not defective}) = \frac{P(\text{acceptable, not defective})}{P(\text{not defective})}$$

Since the transistor will be both acceptable and not defective if it is acceptable.

$$P(\text{acceptable} | \text{not defective}) = \frac{P(\text{acceptable})}{P(\text{not defective})} = \frac{25/40}{35/40} = \frac{5}{7}$$

Essentials of Data Science With R Software - 1

Probability and Statistical Inference

Probability Theory

:::

Lecture 18

Multiplication Theorem of Probability

Shalabh

Department of Mathematics and Statistics

Indian Institute of Technology Kanpur

Multiplication Theorem of Probability:

For two arbitrary events A and B , the following holds:

$$P(A \cap B) = P(A|B)P(B) = P(B|A)P(A).$$

This theorem does not require that $P(A) > 0$ and $P(B) > 0$.

A generalization of this result provides an expression for the probability of the intersection of an arbitrary number of events.

Assume that A_1, A_2, \dots, A_m are events

$$P(A_1 A_2 \dots A_m) = P(A_1) P(A_2|A_1) P(A_3|A_1 A_2) \dots P(A_m|A_1 A_2 \dots A_{m-1})$$

Multiplication Theorem of Probability: Example 1

A student figures that there is a 30 percent chance that he will be selected in the cricket team. If it does, he has 60 percent certain that he will be selected as Captain of the team. What is the probability that the student will be the captain in the selected team?

Solution:

Let T : Event that the student will be selected in the team.

C : Event that the student will be made the captain,

then the desired probability is $P(TC)$.

$$P(TC) = P(T)P(C|T) = (.3)(.6) = 0.18$$

So there is an 18% chance that the student will be the captain.

Multiplication Theorem of Probability: Example 2

A student is undecided as to whether to take a French course or a chemistry course. He estimates that his probability of receiving an A grade would be $\frac{1}{2}$ in a French course, and $\frac{2}{3}$ in a chemistry course. If he decides to base his decision on the flip of a fair coin, what is the probability that he gets an A in chemistry?

Solution:

Let C : event that student takes chemistry and

A : Event that he receives an A in whatever course he takes,

then the desired probability is $P(CA)$. This is calculated as follows:

$$P(CA) = P(C)P(A | C) = \frac{1}{2} \times \frac{2}{3} = \frac{1}{3}.$$

Multiplication Theorem of Probability: Example 3

An ordinary deck of 52 playing cards is randomly divided into 4 piles of 13 cards each. Compute the probability that each pile has exactly 1 ace.

Solution: Define events E_i , $i = 1, 2, 3, 4$ as follows:

E_1 = Event that the ace of spades is in any one of the piles.

E_2 = Event that the ace of spades and the ace of hearts are in different piles

E_3 = Event that the aces of spades, hearts, and diamonds are all in different piles

E_4 = Event that all 4 aces are in different piles

Multiplication Theorem of Probability: Example 3

The probability desired is $P(E_1E_2E_3E_4)$ and by the multiplication rule

$$P(E_1E_2E_3E_4) = P(E_1) \cdot P(E_2 | E_1) \cdot P(E_3 | E_1 E_2) \cdot P(E_4 | E_1 E_2 E_3)$$

- Now $P(E_1) = 1$ since E_1 is the sample space Ω .
- $P(E_2 | E_1) = \frac{39}{51}$ since the pile containing the ace of spades will receive 12 of the remaining 51 cards.
- $P(E_3 | E_1 E_2) = \frac{26}{50}$ since the piles containing the aces of spades and hearts will receive 24 of the remaining 50 cards; and finally,
- $P(E_4 | E_1 E_2 E_3) = \frac{13}{49}$

Multiplication Theorem of Probability: Example 3

Therefore, we obtain that the probability that each pile has exactly 1 ace is

$$P(E_1 E_2 E_3 E_4) = \frac{39}{51} \times \frac{26}{50} \times \frac{13}{49} = 0.105 \text{ (Approximately)}$$

There is approximately a 10.5 percent chance that each pile will contain an ace.

Essentials of Data Science With R Software - 1

Probability and Statistical Inference

Probability Theory

:::

Lecture 19

Bayes' Theorem

Shalabh

Department of Mathematics and Statistics

Indian Institute of Technology Kanpur

Law of Total Probability:

Assume that A_1, A_2, \dots, A_m are events such that

- $A_1 \cup A_2 \cup \dots \cup A_m = \Omega$,
- $A_i \cap A_j = \phi$ (pairwise disjoint) for all $i \neq j = 1, 2, \dots, m$, and
- $P(A_i) > 0$ for all i ,

then the probability of an event B can be calculated as

$$P(B) = \sum_{i=1}^m P(B|A_i)P(A_i)$$

Bayes' Theorem:

Bayes' Theorem gives a connection between $P(A | B)$ and $P(B | A)$.

Consider an example to understand the importance of prior probabilities and Bayes' theorem.

Claim: A blood test for checking the presence/absence of a rare disease is developed with following probabilities:

Events A : Outcome of test is positive

Event D : Person has disease

Bayes' Theorem:

$$P(\text{Person has disease and test is positive}) = P(A | D) = 0.999$$

$$P(\text{Person don't has disease and test is negative}) = P(\bar{A} | \bar{D}) = 0.999$$

Seems to be a good test for a naive person.

Select a person and make a test.

The probability that the person has a disease = $P(D)$

Usually, $P(D)$ is small, say $P(D) = 0.0001$.

Bayes' Theorem:

We want to know whether the test is good or bad.

$$P(\mathbf{D}|\mathbf{A}) = \frac{P(\mathbf{D})P(\mathbf{A}|\mathbf{D})}{P(\mathbf{D})P(\mathbf{A}|\mathbf{D}) + P(\bar{\mathbf{D}})P(\mathbf{A}|\bar{\mathbf{D}})}$$

$$\begin{aligned} &= \frac{0.0001 \times 0.999}{0.0001 \times 0.999 + (1 - 0.0001) \times (1 - 0.999)} \\ &= 0.091 \text{ (Not so reliable, too small)} \end{aligned}$$

So don't rely only on prior probabilities but also look for posterior probabilities.

Bayes' Theorem:

Bayes' Theorem gives a connection between $P(A|B)$ and $P(B|A)$.

For events A and B with $P(A) > 0$ and $P(B) > 0$, we get

$$P(A|B) = \frac{P(AB)}{P(B)} = \frac{P(AB)}{P(A)} \frac{P(A)}{P(B)}$$

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Bayes' Theorem:

Assume that A_1, A_2, \dots, A_m are events such that

- $A_1 \cup A_2 \cup \dots \cup A_m = \Omega$,
- $A_i \cap A_j = \phi$ (pairwise disjoint) for all $i \neq j = 1, 2, \dots, m$
- $P(A_i) > 0$ for all i , and
- B is another event than A , then

$$P(A_j|B) = \frac{P(B|A_j)P(A_j)}{\sum_{i=1}^m P(B|A_i)P(A_i)}$$

is known as Bayes' formula (English philosopher Thomas Bayes).

$P(A_i)$: Prior probabilities

$P(A_i|B)$: Posterior probabilities

$P(B|A_i)$: Model probabilities

Bayes' Theorem: Example 1

Suppose someone rents books from two different libraries.

Sometimes it happens that the book is defective due to missing pages.

We consider the following events:

A_i ($i = 1, 2$): “the book is issued from library i ”.

Further let B denote the event that the book is available and is not defective.

Assume we know that $P(A_1) = 0.6$ and $P(A_2) = 0.4$ and $P(B|A_1) = 0.95$, $P(B|A_2) = 0.75$ and we are interested in the probability that the rented book from the library is not defective.

Bayes' Theorem: Example 1

We can then apply the law of total probability and get

$$P(B) = P(B|A_1)P(A_1) + P(B|A_2)P(A_2) = 0.6 \times 0.95 + 0.4 \times 0.75 = 0.87.$$

We may also be interested in the probability that the book was issued from the library 1 *and* is not defective which is

$$P(B \cap A_1) = P(B|A_1)P(A_1) = 0.95 \times 0.6 = 0.57.$$

Now suppose we have a non-defective book issued. What is the probability that it is issued from library 1?

This is obtained as follows:

$$P(A_1|B) = \frac{P(A_1 \cap B)}{P(B)} = \frac{0.57}{0.87} = 0.6552.$$

Bayes' Theorem: Example 1

Now assume we have a defective book, i.e. \bar{B} occurs.

The probability that a book is defective given that it is from library 1 is $P(\bar{B} | A_1) = 0.05$.

Similarly, $P(\bar{B} | A_2) = 0.25$ for library 2.

We can now calculate the conditional probability that the book is issued from library 1 given that it is defective as follows.

Bayes' Theorem: Example 1

We can now calculate the conditional probability that the book is issued from library 1 given that it is defective:

$$\begin{aligned} P(A_1|\bar{B}) &= \frac{P(\bar{B}|A_1)P(A_1)}{P(\bar{B}|A_1)P(A_1) + P(\bar{B}|A_2)P(A_2)} \\ &= \frac{0.05 \times 0.6}{0.05 \times 0.6 + 0.25 \times 0.4} = 0.2308 \end{aligned}$$

The result about $P(\bar{B})$ used can also be directly obtained by using

$$P(\bar{B}) = 1 - P(B) = 1 - 0.87 = 0.13.$$

Bayes' Theorem: Example 2

At a certain stage of a criminal investigation, the inspector in charge is 60 % convinced of the guilt of a certain suspect.

Suppose now that a *new* piece of evidence that shows that the criminal has a certain characteristic (such as left-handedness, baldness, brown hair, etc.) is uncovered.

If 20 percent of the population possesses this characteristic, how certain of the guilt of the suspect should the inspector now be if it turns out that the suspect is among this group?

Bayes' Theorem: Example 2

Let

G : Event that the suspect is guilty and

C : Event that he possesses the characteristic of the criminal,

we have

$$P(G|C) = \frac{P(GC)}{P(C)}$$

Now

$$P(GC) = P(G)P(C|G) = (.6)(1) = 0.6$$

We have supposed that the probability of the suspect having the characteristic if he is, in fact, innocent is equal to 0.2.

Bayes' Theorem: Example 2

To compute the probability that the suspect has the characteristic, we condition on whether or not he is guilty, we find

$$\begin{aligned} P(C) &= P(C|G)P(G) + P(C|\bar{G})P(\bar{G}) \\ &= (1)(0.6) + (0.2)(0.4) = 0.68 \end{aligned}$$

Hence

$$P(G|C) = \frac{P(GC)}{P(C|G)P(G) + P(C|\bar{G})P(\bar{G})} = \frac{0.60}{0.68} = 0.882$$

and so the inspector should now be 88% certain of the guilt of the suspect.

Bayes' Theorem: Example 3

A plane is missing and it is presumed that it was equally likely to have gone down in any of three possible regions.

The constants p_i are called *overlook probabilities* because they represent the probability of overlooking the plane; they are generally attributable to the geographical and environmental conditions of the regions.

Let $1 - p_i$: Probability the plane will be found upon a search of the i^{th} region when the plane is, in fact, in that region, $i = 1, 2, 3$.

What is the conditional probability that the plane is in the i^{th} region, given that a search of region 1 is unsuccessful, $i = 1, 2, 3$?

Bayes' Theorem: Example 3

Let R_i , $i = 1, 2, 3$, be the event that the plane is in region i ; and let E be the event that a search of region 1 is unsuccessful.

From Bayes' formula, we obtain

$$\begin{aligned} P(R_1|E) &= \frac{P(ER_1)}{P(E)} \\ &= \frac{P(ER_1)}{P(E|R_1)P(R_1)+P(E|R_2)P(R_2)+P(E|R_3)P(R_3)} \\ &= \frac{(p_1)(1/3)}{(p_1)(1/3)+(1)(1/3)+(1)(1/3)} = \frac{p_1}{p_1 + 2} \end{aligned}$$

Bayes' Theorem: Example 3

For $j = 2, 3$,

$$\begin{aligned} P(R_j|E) &= \frac{P(ER_j)}{P(E)} \\ &= \frac{P(ER_j)}{P(E|R_1)P(R_1)+P(E|R_2)P(R_2)+P(E|R_3)P(R_3)} \\ &= \frac{(1)(1/3)}{(p_1)(1/3)+(1)(1/3)+(1)(1/3)} = \frac{1}{p_1 + 2}. \end{aligned}$$

Thus, for instance, if $p_1 = 0.4$, then the conditional probability that the plane is in region 1 given that a search of that region did not uncover it is $1/6$.

Bayes' Theorem: Example 4

In answering a question on a multiple-choice test, a student either knows the answer or guesses.

Let p be the probability that the student knows the answer and $1 - p$ the probability that the student guesses.

Assume that a student who guesses at the answer will be correct with probability $\frac{1}{m}$, where m is the number of multiple-choice alternatives.

What is the conditional probability that a student knew the answer to a question, given that he or she answered it correctly?

Bayes' Theorem: Example 4

Solution: Let

C : Events that the student answers the question correctly and

K : Event that he or she actually knows the answer.

Now

$$\begin{aligned} P(K|C) &= \frac{P(KC)}{P(C)} = \frac{P(C|K)P(K)}{P(C|K)P(K)+P(C|\bar{K})P(\bar{K})} \\ &= \frac{p}{p + (\frac{1}{m})(1 - p)} = \frac{mp}{1 + (m - 1)p} \end{aligned}$$

For example, if $m = 4$, $p = 0.5$, then the probability that a student knew the answer to a question he or she correctly answered is $4/5$.

Bayes' Theorem: Example 5

A laboratory blood test is 95 percent effective in detecting a certain disease when it is, in fact, present.

However, the test also yields a "false positive" result for 1% of the healthy persons tested. (i.e., if a healthy person is tested, then, with probability .01, the test result will imply he or she has the disease.)

If 0.5% of the population actually has the disease, what is the probability a person has the disease given that the test result is positive?

Bayes' Theorem: Example 5

Solution: Let D : Event that the tested person has the disease and

E : Event that the test result is positive.

The desired probability $P(D | E)$ is obtained as follows:

$$\begin{aligned} P(D|E) &= \frac{P(DE)}{P(E)} = \frac{P(E|D)P(D)}{P(E|D)P(D)+P(E|\bar{D})P(\bar{D})} \\ &= \frac{(0.95)(0.005)}{(0.95)(0.005)+(0.01)(0.995)} = \frac{95}{294} = 0.323(\text{Approx.}) \end{aligned}$$

Thus only 32% of those persons whose test results are positive actually have the disease.

Surprised!! As we expected this figure to be much higher, since the blood test seems to be a good one.