Lingnan University 嶺南大學 香港 Hong Kong

School of Data Science 數據科學學院

CDS 533
Statistics for Data Science

Instructor: Lisha Yu

Division of Artificial Intelligence

School of Data Science

Lingnan University

Fall 2024

Let’s Know Each Other

Course Outline

(Sep 2nd - Nov 25th , NO CLASS ON OCT 7th)

This Course: Fundamental Knowledge

Hands-on Examples

The World of Statistics

We Are Here to Support

Instructor: Dr. Lisha Yu

Time: Monday 13:30 PM - 17:00 PM

Office: AD105/3

Phone: (852)2616-8737

E-mail: lishayu@ln.edu.hk

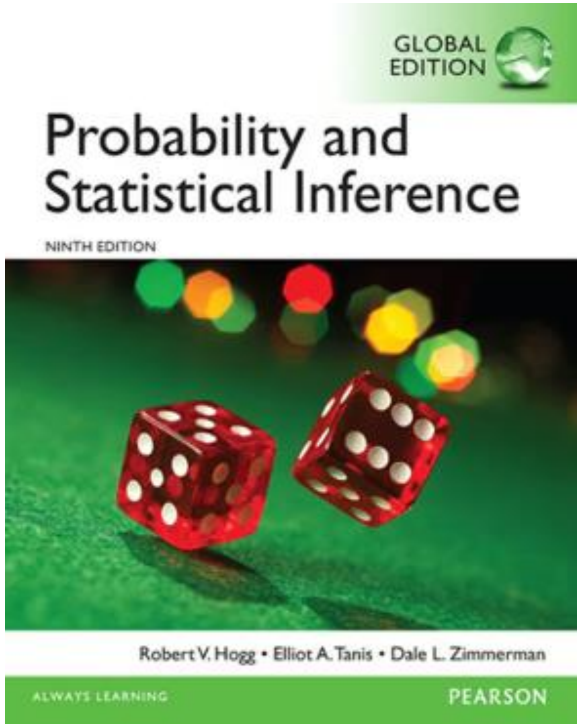
Office Hours: T 11:00-12:30 14:00-15:00 THUR 15:00-16:30

TA: Yugao HE

E-mail: yugaohe@ln.edu.hk

Books

(Basic)



GLOBAL EDITION

Probability and Statistical Inference

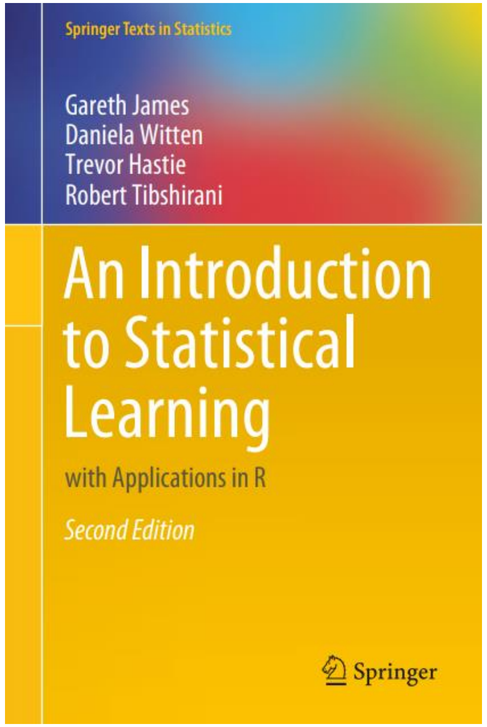
NINTH EDITION

Robert V. Hogg • Elliot A. Tanis • Dale L. Zimmerman

ALWAYS LEARNING

PEARSON

(Basic)



Springer Texts in Statistics

Gareth James
Daniela Witten
Trevor Hastie
Robert Tibshirani

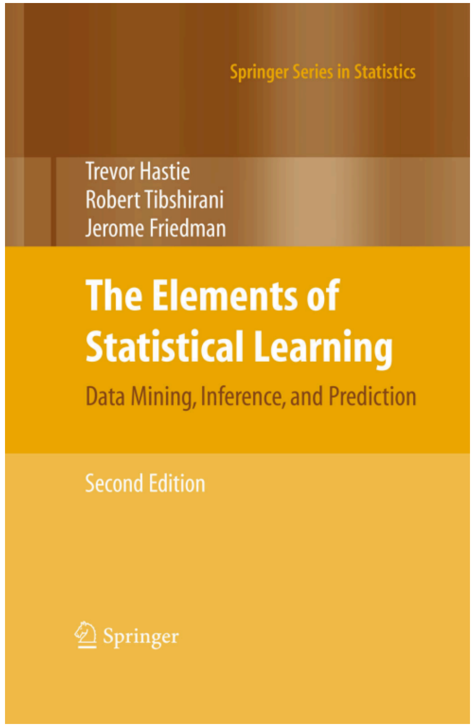
An Introduction to Statistical Learning

with Applications in R

Second Edition

Springer

(Advanced)



Springer Series in Statistics

Trevor Hastie
Robert Tibshirani
Jerome Friedman

The Elements of Statistical Learning

Data Mining, Inference, and Prediction

Second Edition

Springer

Tentative Schedule

Week	Date	Lecture	Material	Notes	
1	2/9	1	Introduction		
2	9/9	2	Sampling EDA		
3	9/16	3	EDA Sampling Distribution	Assignment 1	
4	9/23	4	Estimation and Hypothesis test		
5	9/30	5	Statistical Experiment and Analyze of Variance	Assignment 2	Due A1
6	10/7	No class			
7	10/14	6	Regression		Due A2
8	10/21	7	Regression	Assignment 3	
9	10/28	8	Classification		
10	11/4	9	Classification		Due A3
11	11/11	10	Resampling	Assignment 4	
12	11/18	11	Time series		
13	11/25	12	Revision		Due A4

For Today’s Graduate, Just One Word: Statistics

- *New York Times*, *2019*

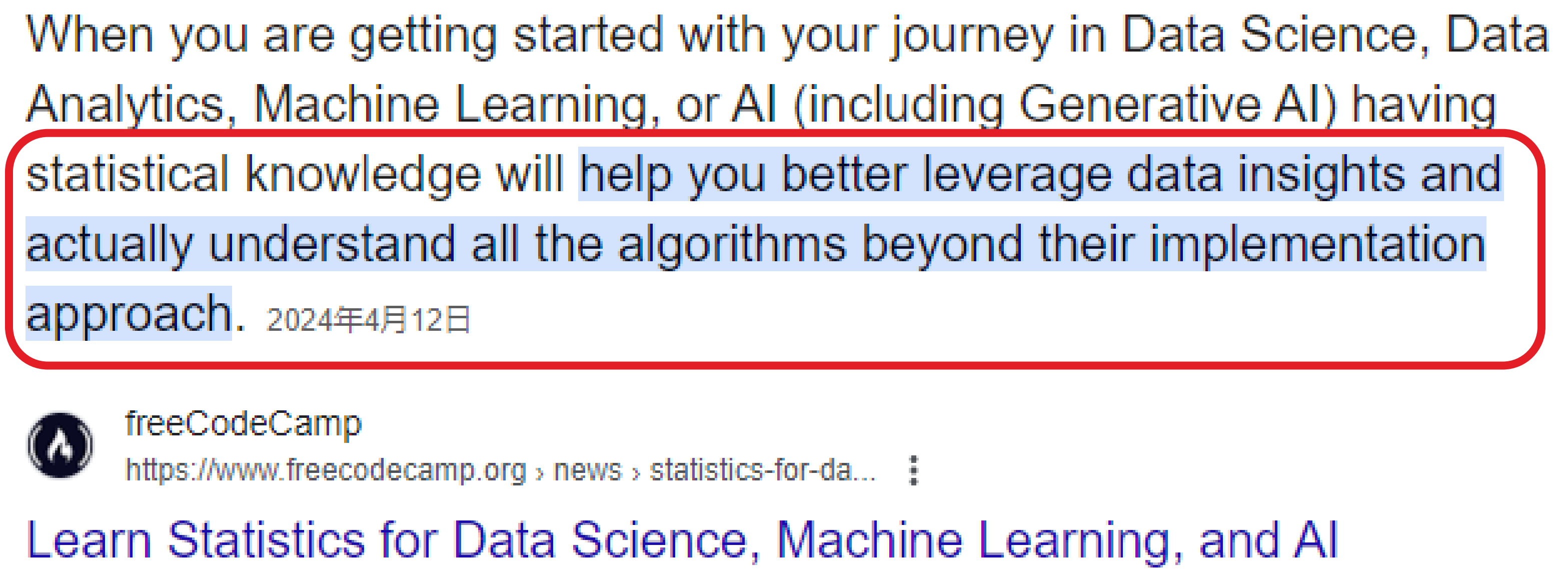
(https://www.nytimes.com/2009/08/06/technology/06stats.html)

Statistics and Data Science

Data Science is one of the most trending buzzwords nowadys.

WHY STATISTICS?

Statistics and Data Science



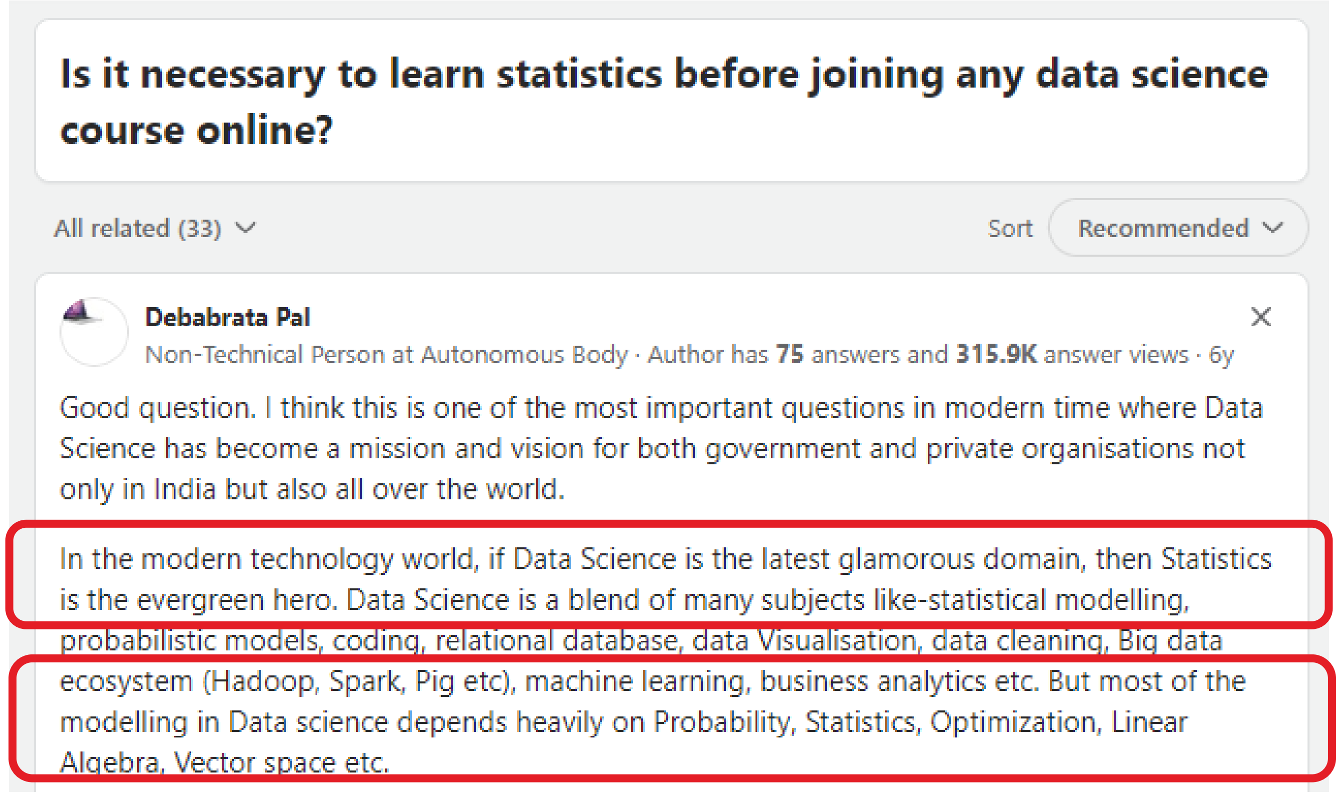
When you are getting started with your journey in Data Science, Data Analytics, Machine Learning, or AI (including Generative AI) having statistical knowledge will help you better leverage data insights and actually understand all the algorithms beyond their implementation approach. 2024年4月12日

freeCodeCamp
https://www.freecodecamp.org › news › statistics-for-da...

Learn Statistics for Data Science, Machine Learning, and AI

Google

Statistics and Data Science



Is it necessary to learn statistics before joining any data science course online?

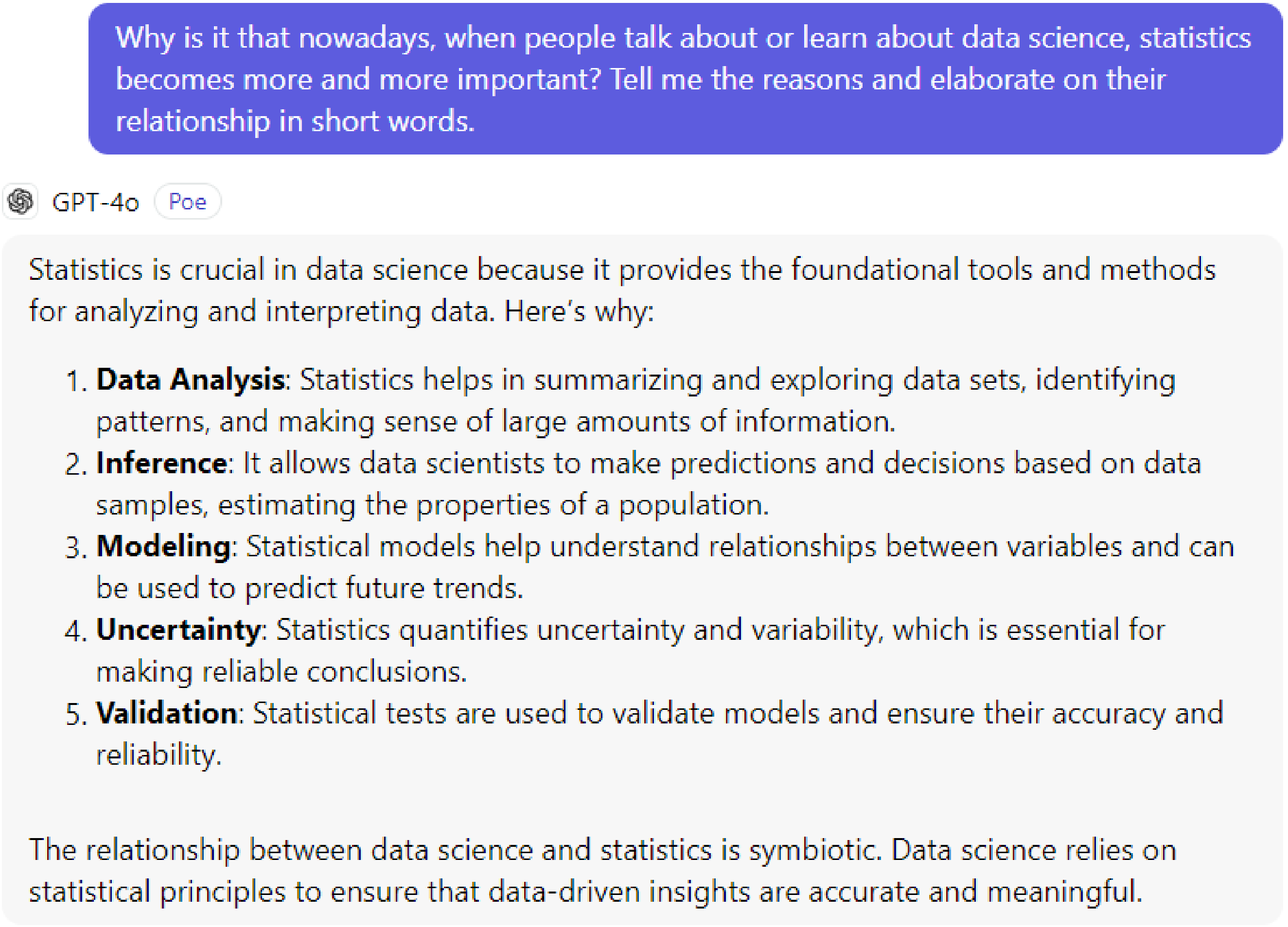
Debabrata Pal
Non-Technical Person at Autonomous Body · Author has **75** answers and **315.9K** answer views · 6y

Good question. I think this is one of the most important questions in modern time where Data Science has become a mission and vision for both government and private organisations not only in India but also all over the world.

In the modern technology world, if Data Science is the latest glamorous domain, then Statistics is the evergreen hero. Data Science is a blend of many subjects like-statistical modelling, probabilistic models, coding, relational database, data Visualisation, data cleaning, Big data ecosystem (Hadoop, Spark, Pig etc), machine learning, business analytics etc. But most of the modelling in Data science depends heavily on Probability, Statistics, Optimization, Linear Algebra, Vector space etc.

Forum

Statistics and Data Science



Why is it that nowadays, when people talk about or learn about data science, statistics becomes more and more important? Tell me the reasons and elaborate on their relationship in short words.

Statistics is crucial in data science because it provides the foundational tools and methods for analyzing and interpreting data. Here's why:

1. **Data Analysis**: Statistics helps in summarizing and exploring data sets, identifying patterns, and making sense of large amounts of information.
2. **Inference**: It allows data scientists to make predictions and decisions based on data samples, estimating the properties of a population.
3. **Modeling**: Statistical models help understand relationships between variables and can be used to predict future trends.
4. **Uncertainty**: Statistics quantifies uncertainty and variability, which is essential for making reliable conclusions.
5. **Validation**: Statistical tests are used to validate models and ensure their accuracy and reliability.

The relationship between data science and statistics is symbiotic. Data science relies on statistical principles to ensure that data-driven insights are accurate and meaningful.

(ChatGPT, 2024)

GPT

Statistics is the backbone of data science.

Tips

Having a good understanding of *Probability and Statistics* is a **must** if you are willing to **become a data science professional.**

Missing the Big Picture

After introductory statistics course

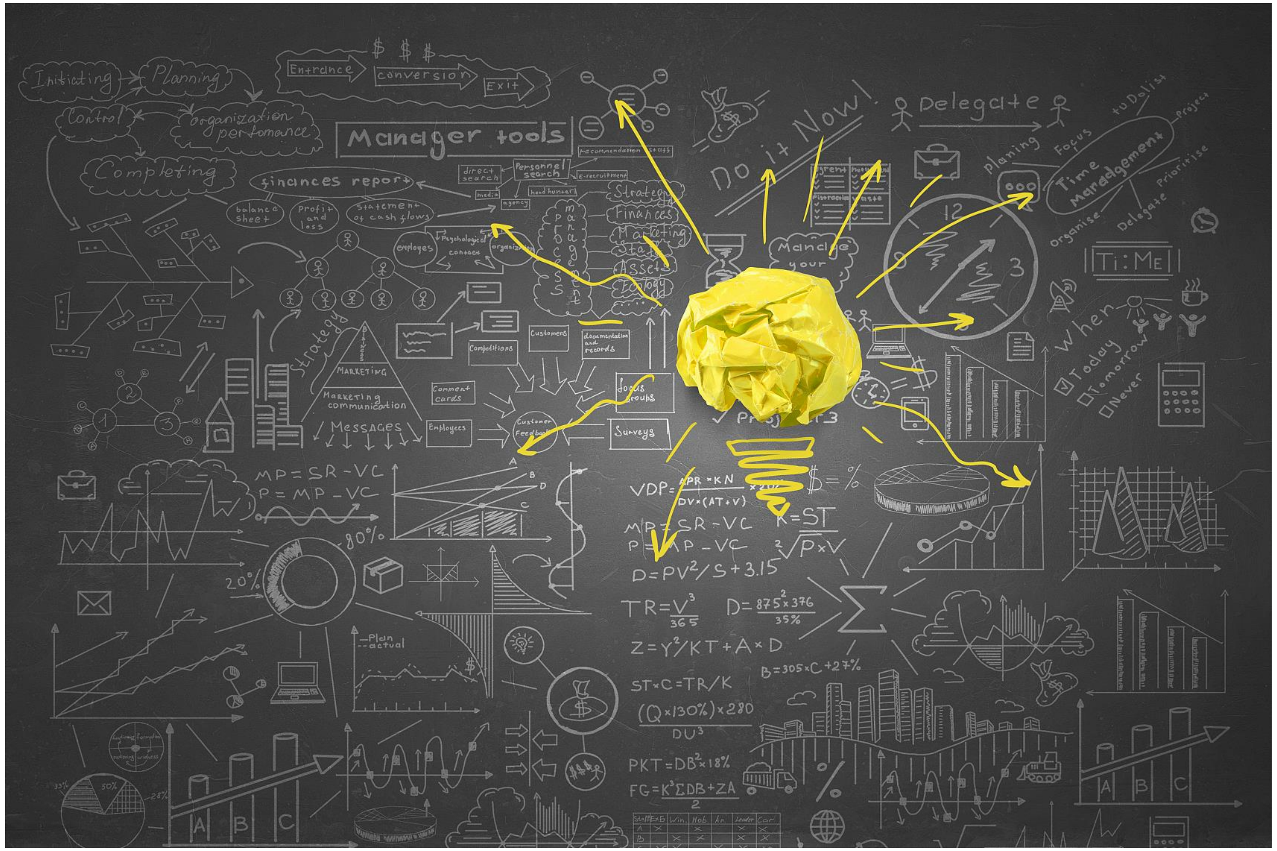
- See statistics as disconnected topics
- Learn normal distribution for the first exam
  - Then forget it
- Learn confidence intervals for the second exam
  - Then forget it

Never see connection

- Some don't link 1.96 in a margin of error with normal distribution

Bigger problem: Don't understand where statistics will be in their lives.
WHY do I care?

Class Philosophy

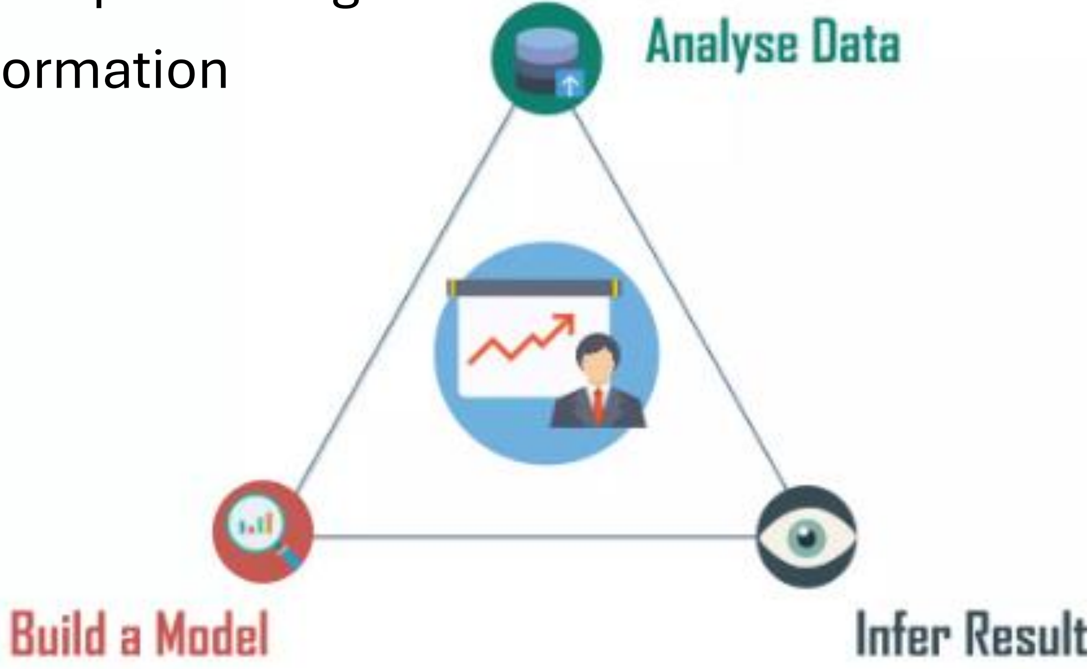


Let statistics inspire you to see the world in a new light
Enlighten your mind with the power of data and discovery

Introduction to Statistics

Statistics is a branch of mathematics dealing with data collection and organization, analysis, interpretation, and presentation.

- ***Collection - the process of obtaining information***
- Organizaiton - ascertaining the manner of presenting the data
- Analysis - the process of extracting information

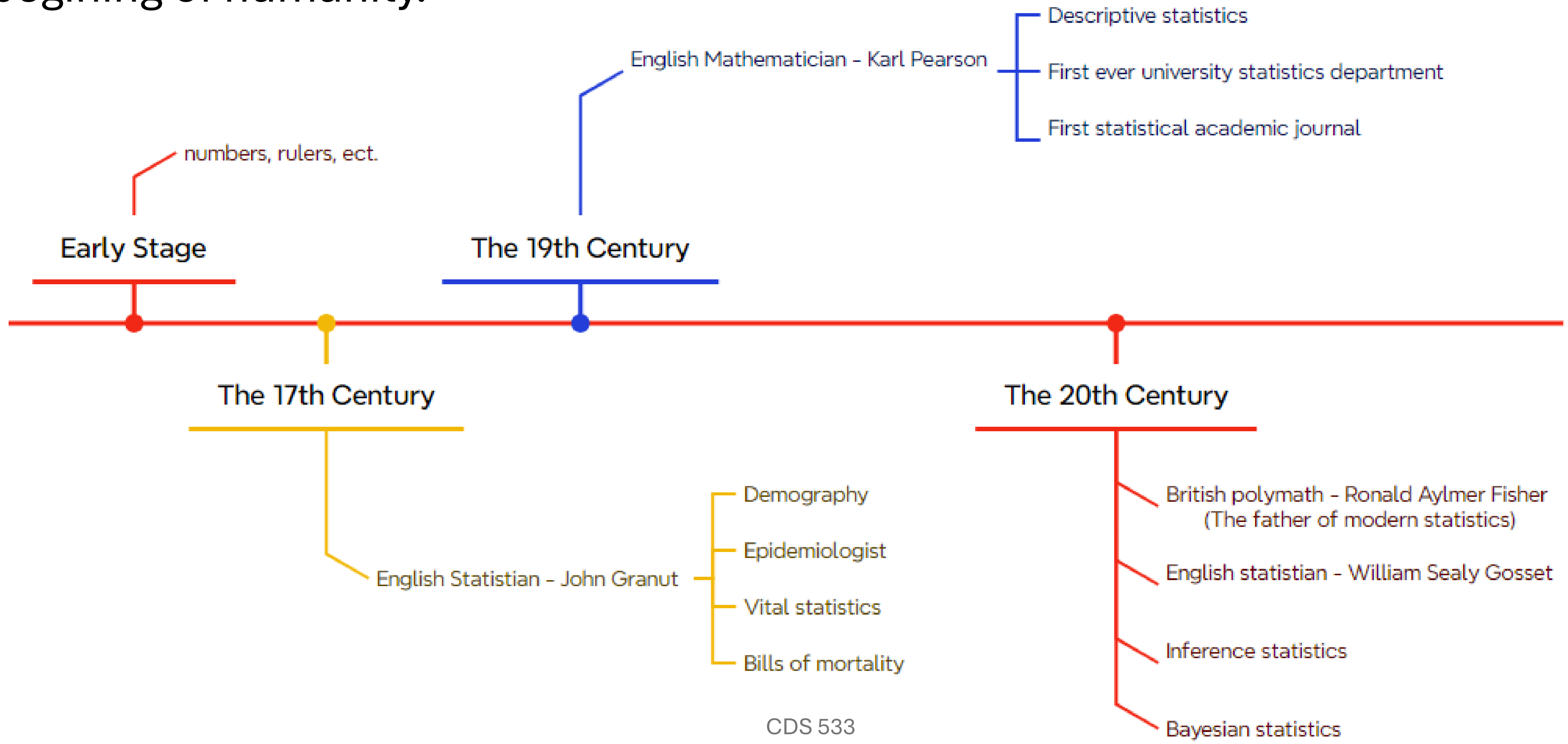


History of Statistics

The Word statistics has been derived from the Latin word “Status” or the Italian word “Statista”; meaning of these words is “Political State.”

Journey of Statistics

The processing of statistical informaiton has a history that extends back to the begining of humanity.


numbers, rulers, ect.
Early Stage
English Mathematician - Karl Pearson
Descriptive statistics
First ever university statistics department
First statistical academic journal
The 19th Century
The 17th Century
English Statistian - John Granut
Demography
Epidemiologist
Vital statistics
Bills of mortality
The 20th Century
British polymath - Ronald Aylmer Fisher (The father of modern statistics)
English statistian - William Sealy Gosset
Inference statistics
Bayesian statistics

What can Statistics do?

Senario 1: Your company has invented a new drug that may cure cancer. How would you conduct a test to confirm the drug's effectivess?

What can Statistics do?

Senario 2: The latest sale data have just come in, and your boss wants you to prepare a report for managment on places where the company could improve the business.What should you look for? What should you not look for?

Importance of Statistics

Statistics is a universal data analysis method applicable across all domains - wherever data exists, statistical techniques can be applied.

- As data has become increasingly valued, statistical methods have expanded into numerous fields in natural and social sciences, evolving into an interdisciplinary system of specialized sub-disciplines.

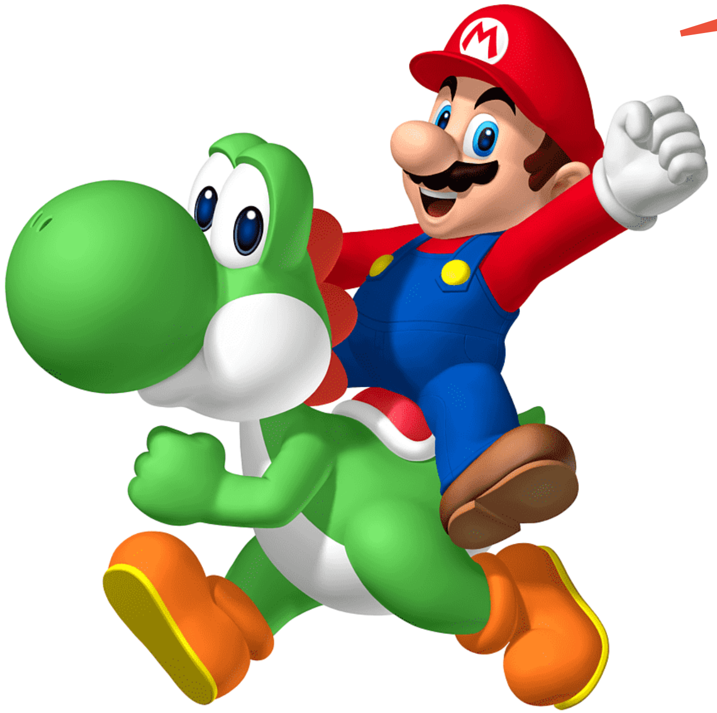
Importance of Statistics

Statistical methods can be used to find answers to questions like...

- What kind and how much data need to be collected?
- How should we organize and summarize the data?
- How can we analyze the data and draw conclusions from it?
- How can we assess the strength of the conclusions and evaluate their uncertainty?
- ...

(Dall-E, 2024)

Application of Statistics



Let me show you some statistical scenarios in my world!

Application of Statistics

Scenario 1: We want to estimate the average time it takes ALL players to complete the "Cloud Kingdom" level.

Scenario 2: The developers believe the new Fire Flower power-up significantly boosts Mario's ability to defeat enemies with fireballs.

Scenario 3: We want to collect player satisfaction with the new game update that introduces Yoshi as a playable character.

Scenario 4: We want to understand the relationship between how long players explore the Mushroom Kingdom and their in-game experience points (XP).

Tips

When you learn statistics, remember that

statistics

is not only

a collection of many

data analysis techniques

it is

An integrated system for

thinking with data

Statistical Thinking

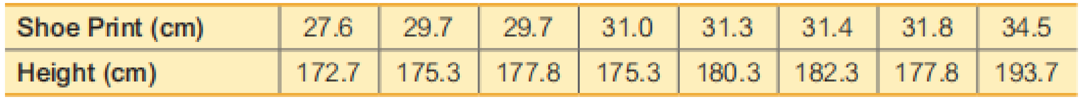
Whether conducting statistical analysis of data that we have collected or analyzing a statistical analysis done by someone else, we should not rely on blind acceptance of mathematical calculation. We should consider these factors:

- Context of the data
- Source of the data
- Sampling method
- Statistical analysis
- Conclusions
- Practical implications

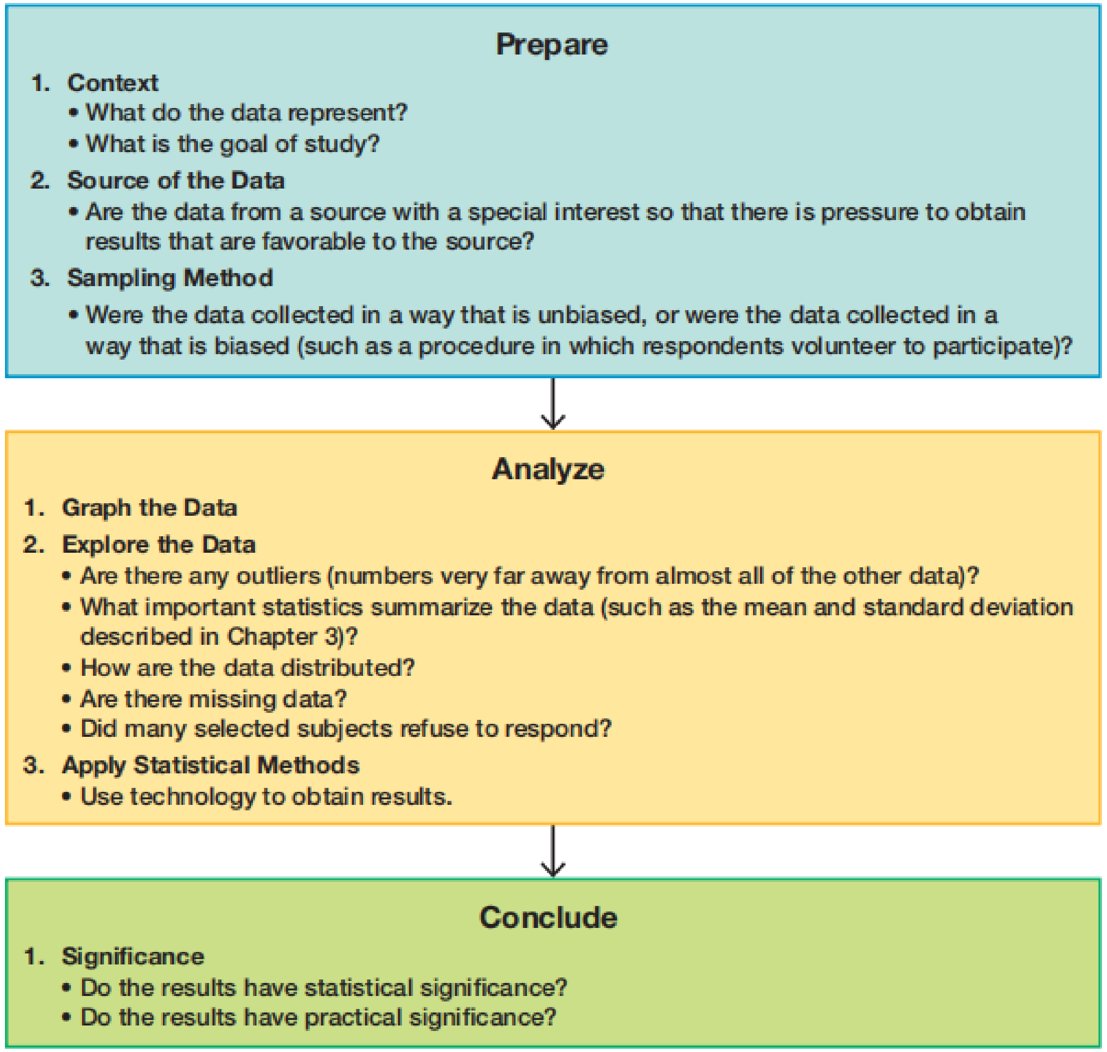
Statistical Thinking

Example

TABLE 1-1 Shoe Print Lengths and Heights of Men



Shoe Print (cm)	27.6	29.7	29.7	31.0	31.3	31.4	31.8	34.5
Height (cm)	172.7	175.3	177.8	175.3	180.3	182.3	177.8	193.7



Prepare

1. **Context**
   - What do the data represent?
   - What is the goal of study?
2. **Source of the Data**
   - Are the data from a source with a special interest so that there is pressure to obtain results that are favorable to the source?
3. **Sampling Method**
   - Were the data collected in a way that is unbiased, or were the data collected in a way that is biased (such as a procedure in which respondents volunteer to participate)?

Analyze

1. **Graph the Data**
2. **Explore the Data**
   - Are there any outliers (numbers very far away from almost all of the other data)?
   - What important statistics summarize the data (such as the mean and standard deviation described in Chapter 3)?
   - How are the data distributed?
   - Are there missing data?
   - Did many selected subjects refuse to respond?
3. **Apply Statistical Methods**
   - Use technology to obtain results.

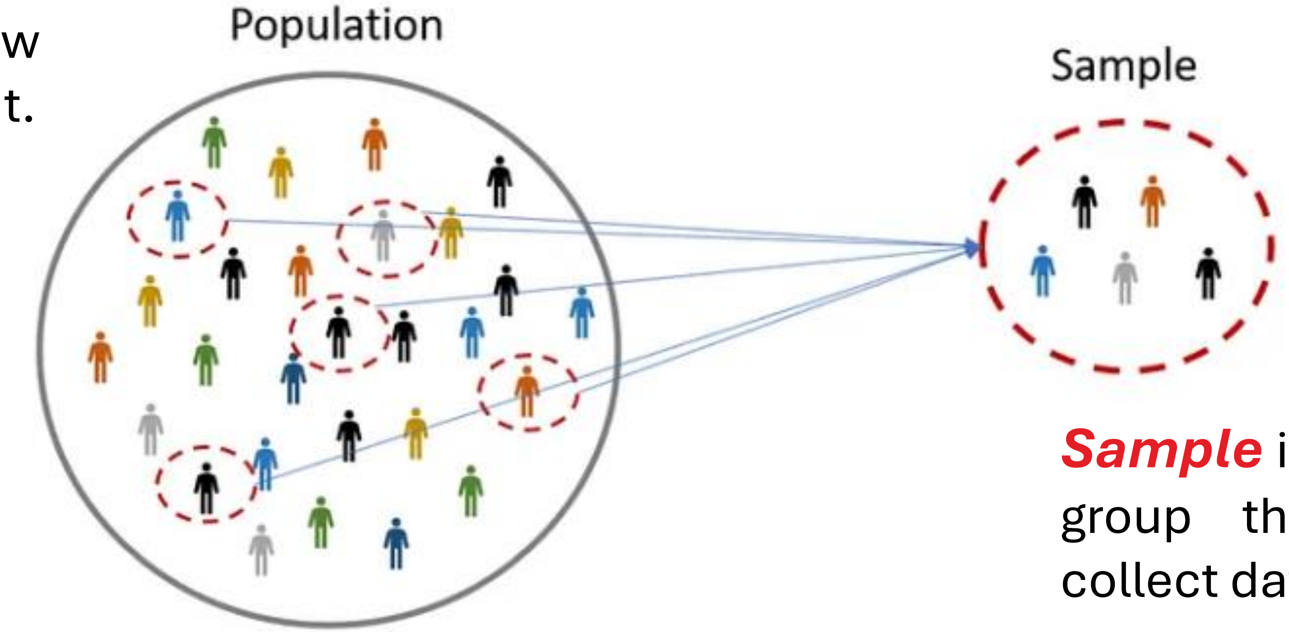
Conclude

1. **Significance**
   - Do the results have statistical significance?
   - Do the results have practical significance?

FIGURE 1-3 **Statistical and Critical Thinking**

Basic Terminologies in Statistics

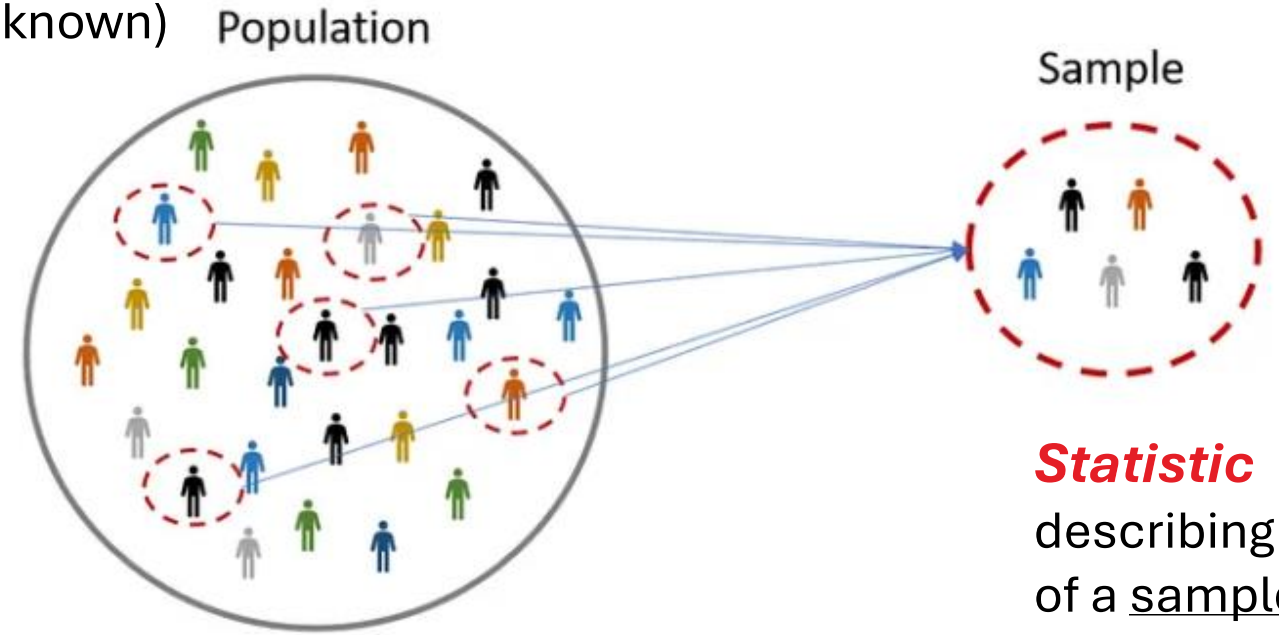
Population is the entire group that you want to draw conclusions about.



Sample is the specific group that you will collect data from.

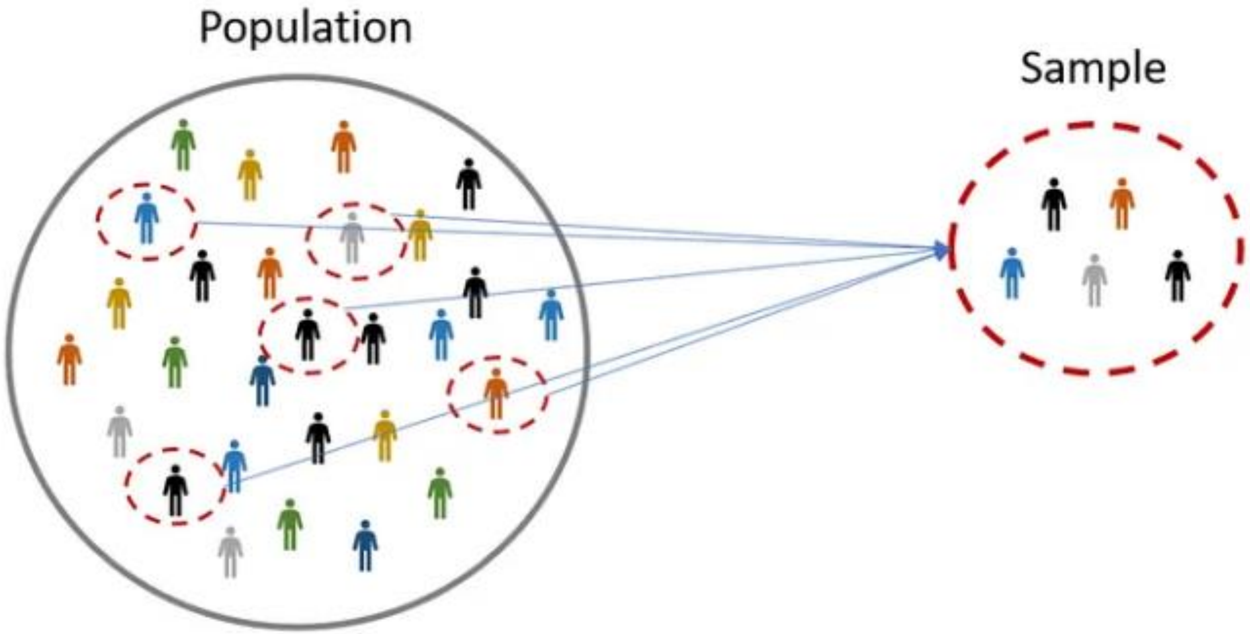
Basic Terminologies in Statistics

Parameter is a numerical measurement describing some characteristic of a population. (unknown)



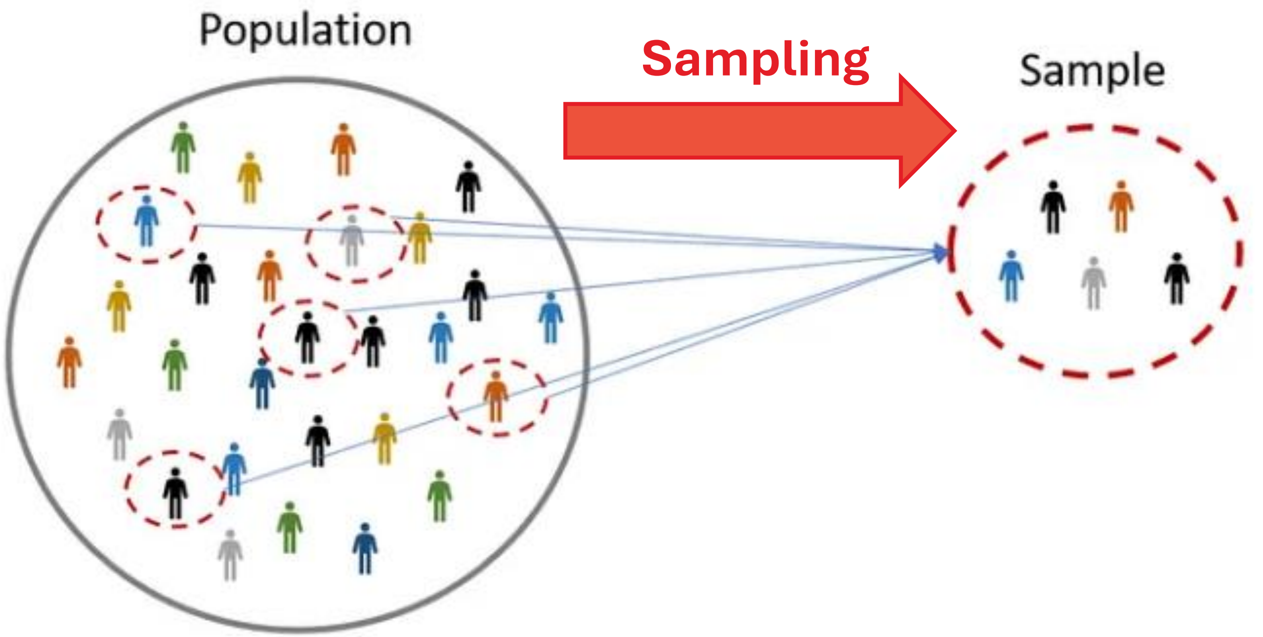
Statistic is a measurement describing some characteristic of a sample. (known)

Basic Terminologies in Statistics



Parameter		Statistics
μ	Mean	$\bar{x}$
σ	Standard deviation	s
π	Proportion	p
N	Size	n

Types of Statistics

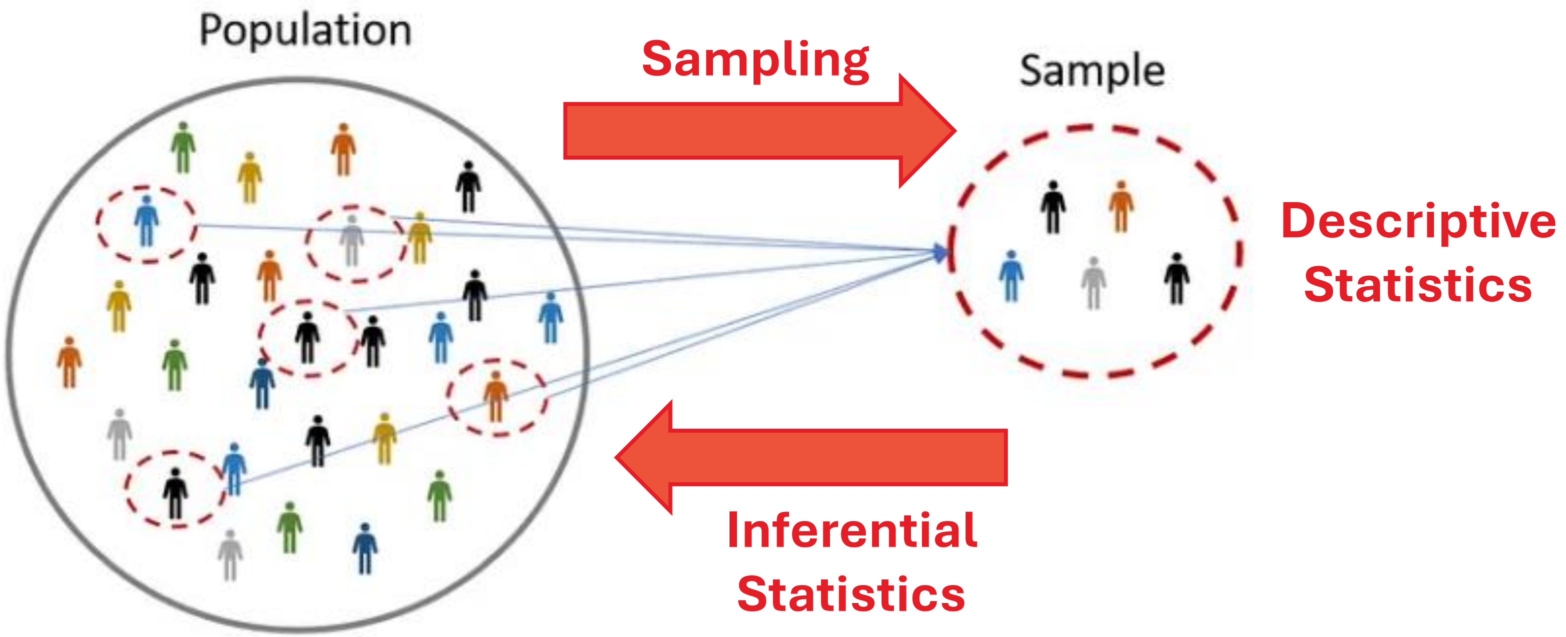


Sampling

Descriptive Statistics

Descriptive statistics is a method used to describe and understand the features of a specific data set by giving short summaries about the sample and measures of the data.

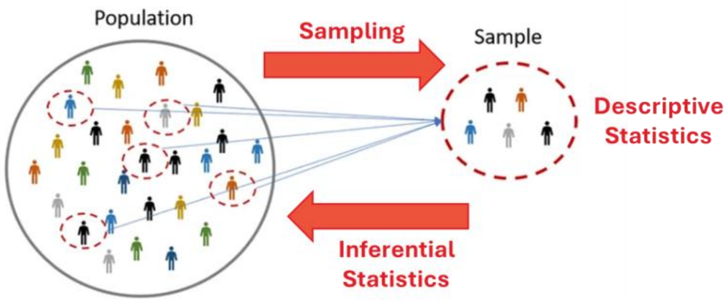
Types of Statistics

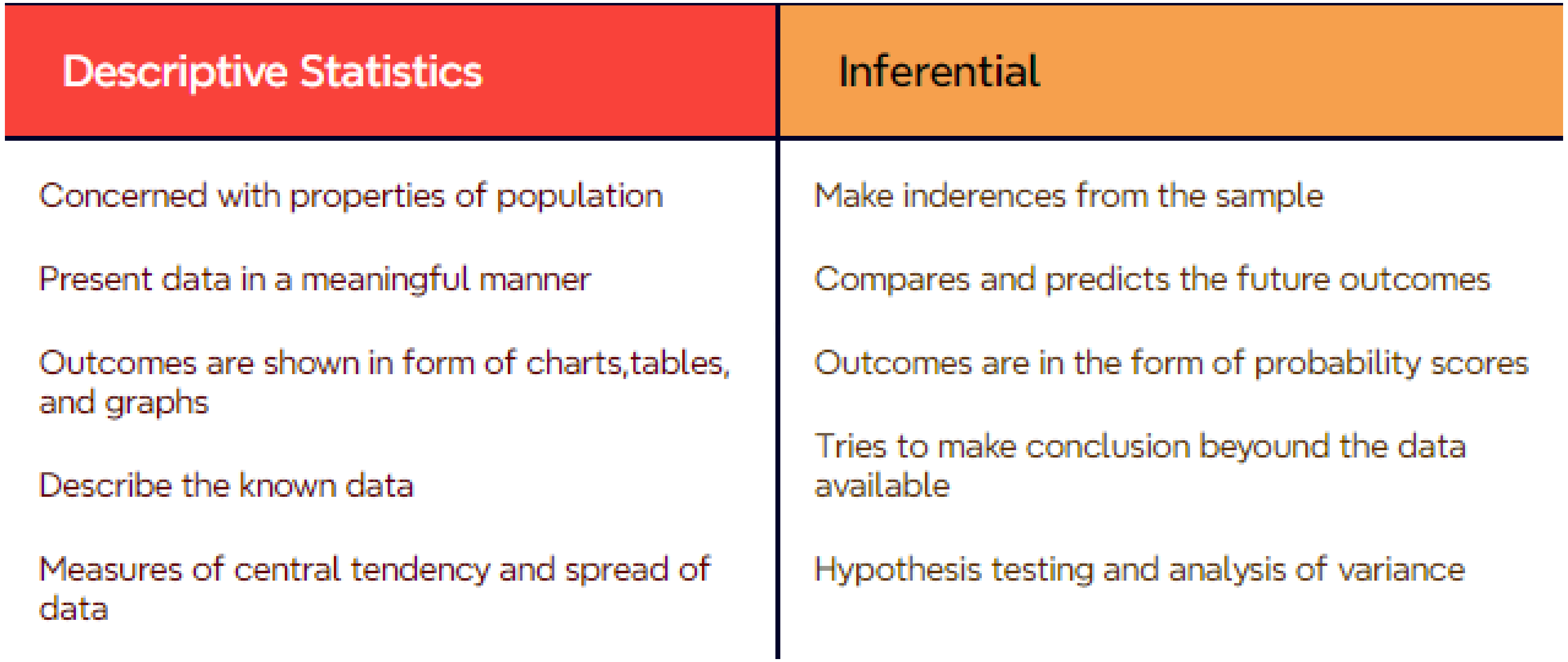


Descriptive statistics is a method used to describe and understand the features of a specific data set by giving short summaries about the sample and measures of the data.

Inferential statistics makes inferences and predictions about a population based on a sample of data taken from the population in question.

Types of Statistics





Descriptive Statistics	Inferential
Concerned with properties of population	Make inderences from the sample
Present data in a meaningful manner	Compares and predicts the future outcomes
Outcomes are shown in form of charts,tables, and graphs	Outcomes are in the form of probability scores
Describe the known data	Tries to make conclusion beyound the data available
Measures of central tendency and spread of data	Hypothesis testing and analysis of variance

Review

Q. Which term describes a numerical measurement that characterizes an entire population?

A. Parameter
B. Estimate
C. Statistic
D. Sample

ANSWER: A

Q. Which of the following best defines inferential statistics?

A. A method for analyzing components of a sample
B. Statistics that only summarizes data without any predictions
C. Statistics that make predictions and inferences about a population
D. An approach to collect data from a non-representative sample

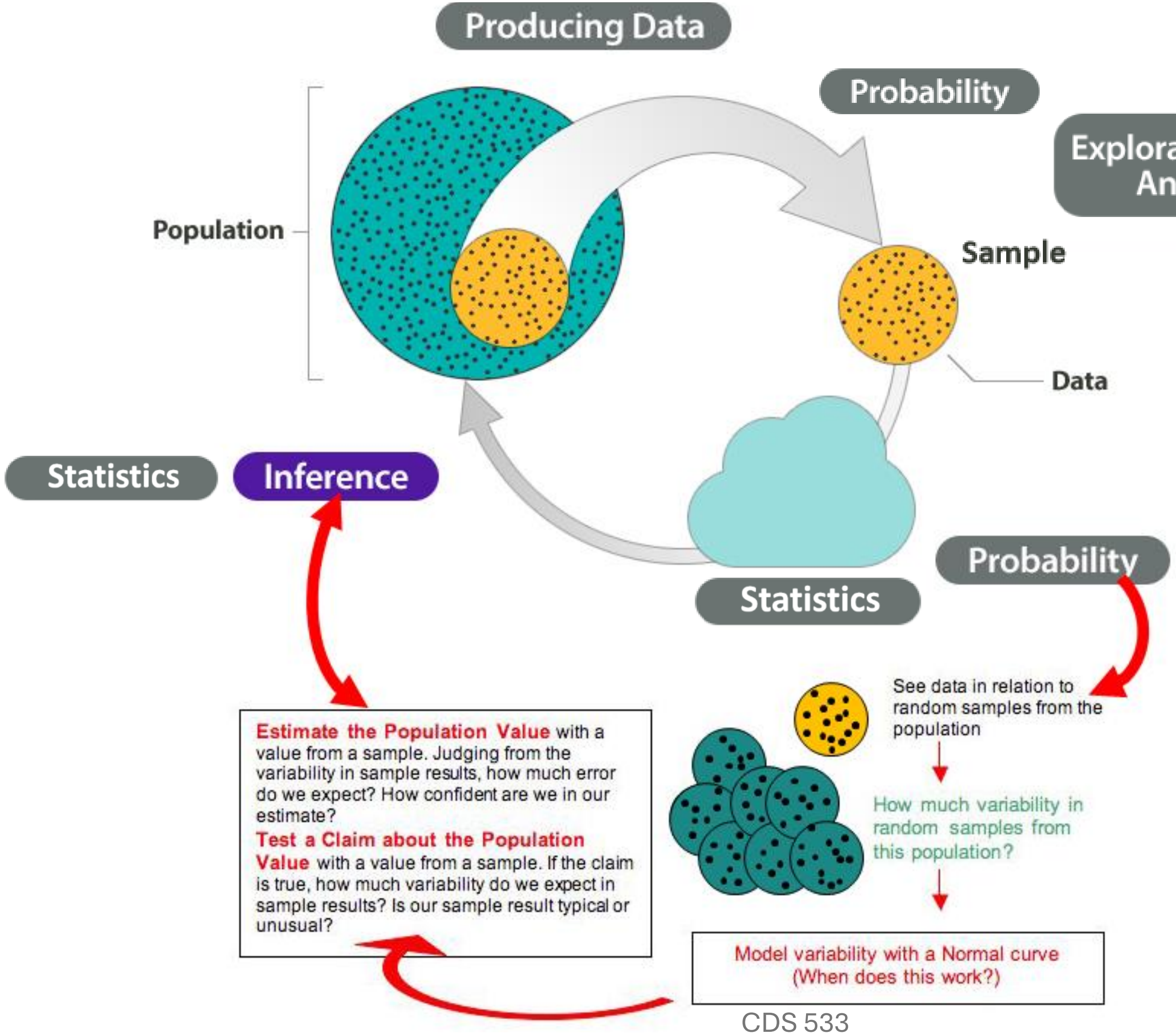
ANSWER: C

Q. Given data on 20 fish caught in a lake, what’s the average weight of all fish in the lake?

A. Descritive
B. Inferential

ANSWER: B

Big Picture of Statistics



Estimate the Population Value with a value from a sample. Judging from the variability in sample results, how much error do we expect? How confident are we in our estimate?
Test a Claim about the Population Value with a value from a sample. If the claim is true, how much variability do we expect in sample results? Is our sample result typical or unusual?

See data in relation to random samples from the population

How much variability in random samples from this population?

Model variability with a Normal curve (When does this work?)

- Sampling
- Exploratory Data Analysis
- Statistical Inference
- Modelling
- Resampling

Lab Time

What is R?

R is an open-source computer language used for data manipulation, statistics, and graphics.

History of R

- 1976 – Bell Labs develops S (S-plus), a language for data analysis
- 1990s – R written and released as open source by (R)oss Ihaka and (R)obert Gentleman
- 1997 – The Comprehensive R Archive Network (CRAN) launched

Till now…

- More than 13 thousand libraries in R are focused on helping data scientists with data analysis, statistical modeling, time series analysis, and machine learning algorithms.

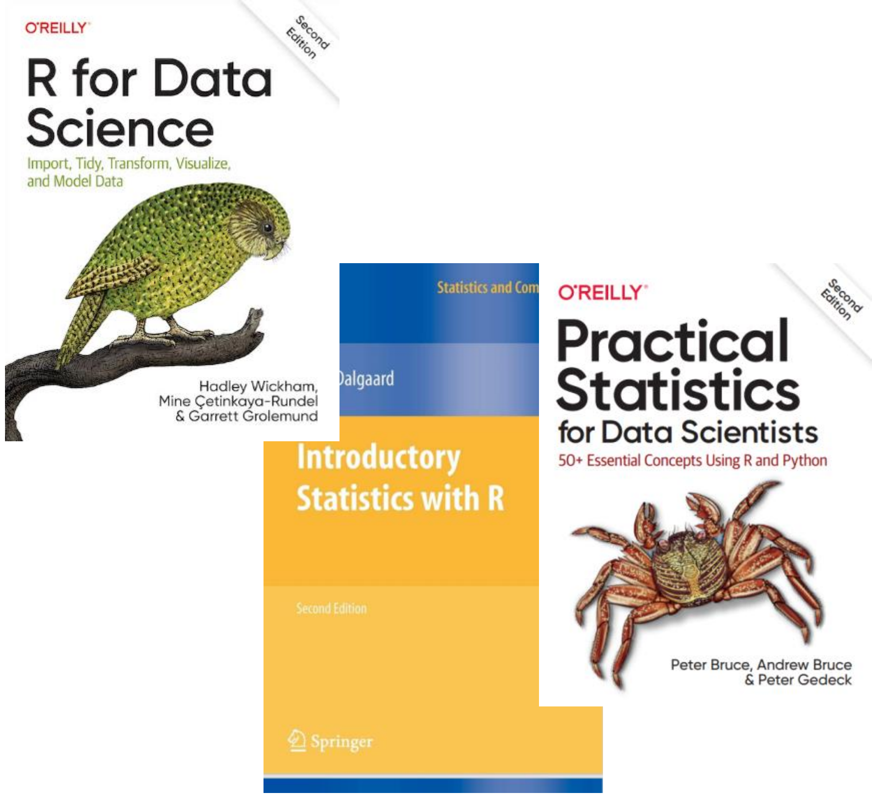
DID YOU KNOW: R's ggolot2 is the most downloaded graphical package in the world

Reasons for Moving to R

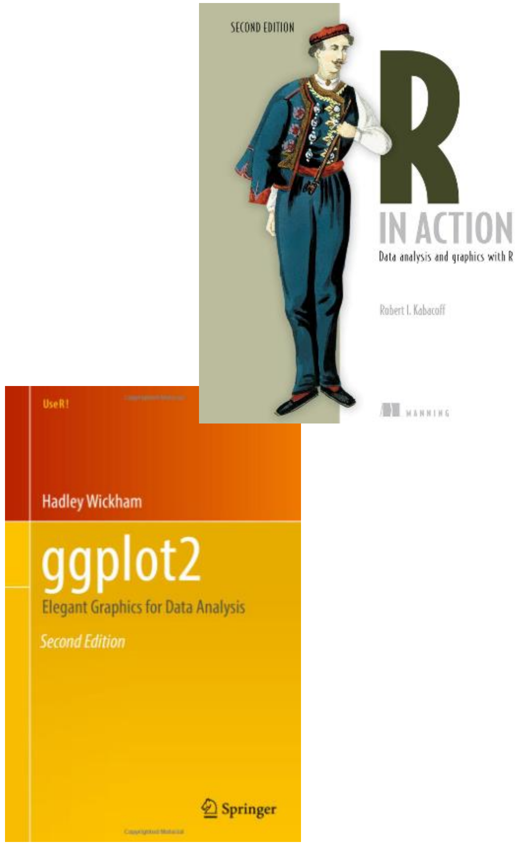
- R is open-source and freely available
- R is cross-platform compatible
  - Runs on multiple platforms (Windows, Unix, MacOS)
  - Adpated with other software (call functions)
- R is powerfully interpreted computer language
- R is highly flexible and evolved with a strong user community
- R has the implementation of high-end statistical methods

What is R?

Basic


O'REILLY
Second Edition
R for Data Science
Import, Tidy, Transform, Visualize, and Model Data
Hadley Wickham, Mine Çetinkaya-Rundel & Garrett Grolemund
Statistics and Com
Dalgaard
Introductory Statistics with R
Second Edition
Springer
O'REILLY
Second Edition
Practical Statistics for Data Scientists
50+ Essential Concepts Using R and Python
Peter Bruce, Andrew Bruce & Peter Gedeck

Handouts/Kaggle/Datacamp
https://rstudio.github.io/cheatsheets
https://r4ds.had.co.nz/index.html



More readings

Learning Curve

The bad news:

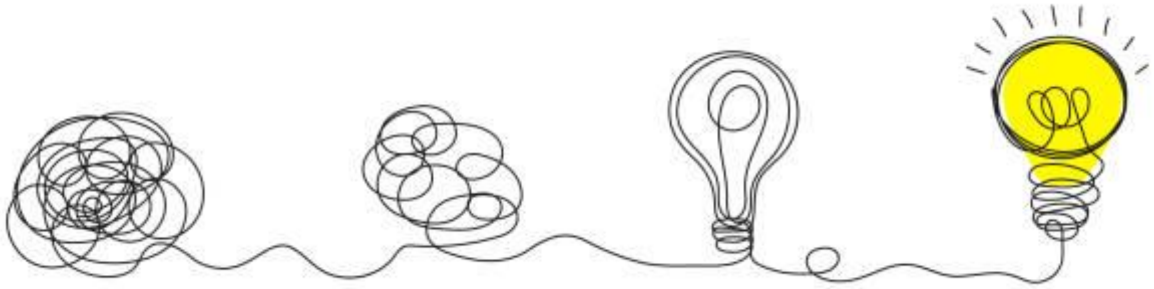
It’s going to be frustrating

Learning Curve

The good news:

Frustrating is typical and temporary

R has a steep learning curve



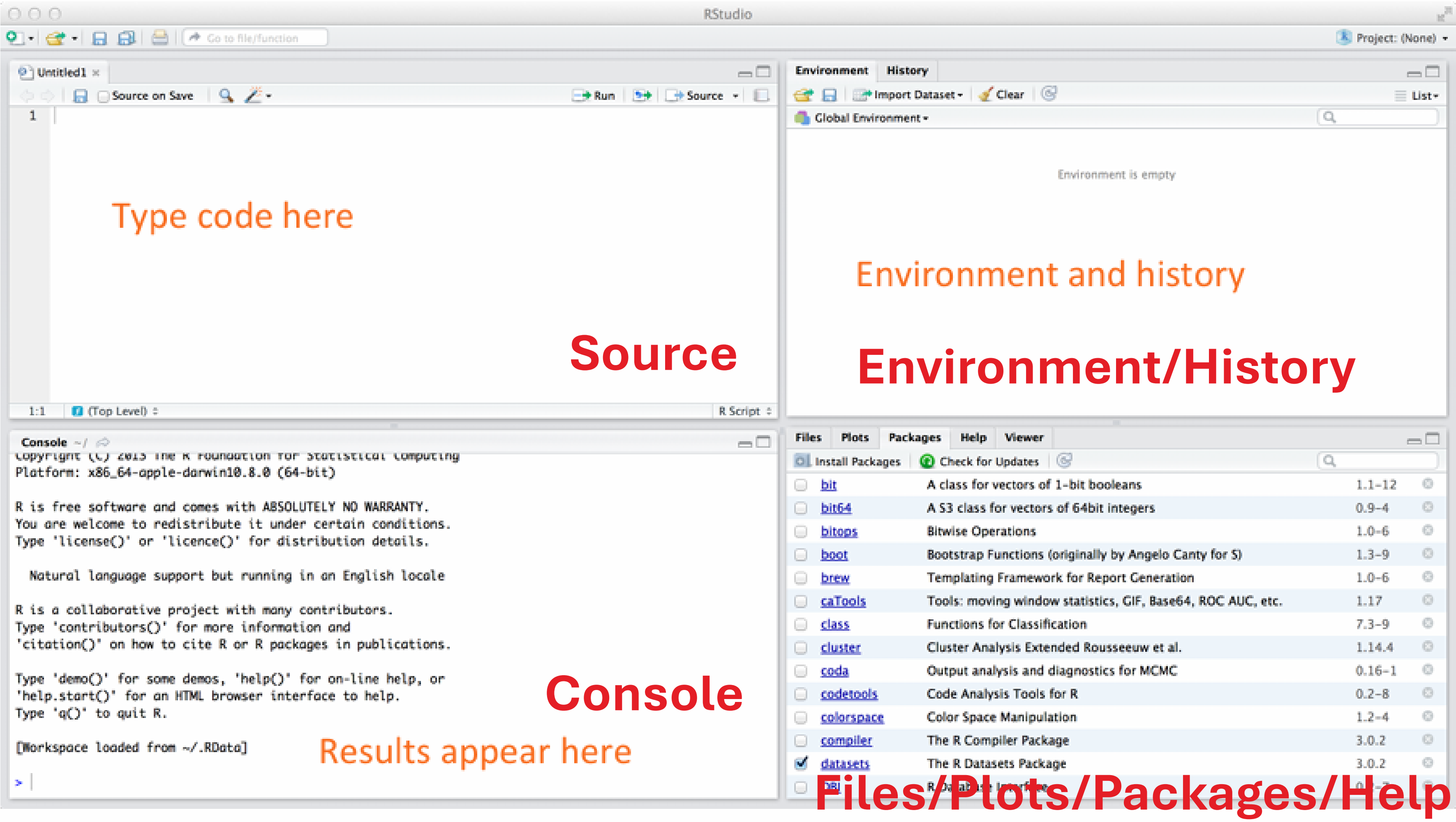
Setting up

1. Go to www.r-project.org and download R
2. Or...Go to https://posit.co/ and download R and Rstudio

Then.....You are all set up to learn R

3. RStudio Cloud: https://rstudio.cloud

RStudio Interface



```
Copyright (C) 2013 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin10.8.0 (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.

[Workspace loaded from ~/.RData]

>
```

□	bit	A class for vectors of 1-bit booleans	1.1-12
□	bit64	A S3 class for vectors of 64bit integers	0.9-4
□	bitops	Bitwise Operations	1.0-6
□	boot	Bootstrap Functions (originally by Angelo Canty for S)	1.3-9
□	brew	Templating Framework for Report Generation	1.0-6
□	caTools	Tools: moving window statistics, GIF, Base64, ROC AUC, etc.	1.17
□	class	Functions for Classification	7.3-9
□	cluster	Cluster Analysis Extended Rousseeuw et al.	1.14.4
□	coda	Output analysis and diagnostics for MCMC	0.16-1
□	codetools	Code Analysis Tools for R	0.2-8
□	colorspace	Color Space Manipulation	1.2-4
□	compiler	The R Compiler Package	3.0.2
☑	datasets	The R Datasets Package	3.0.2
□	DBI	R Database Interface	[illegible]

Type code here

Source

Environment and history

Environment/History

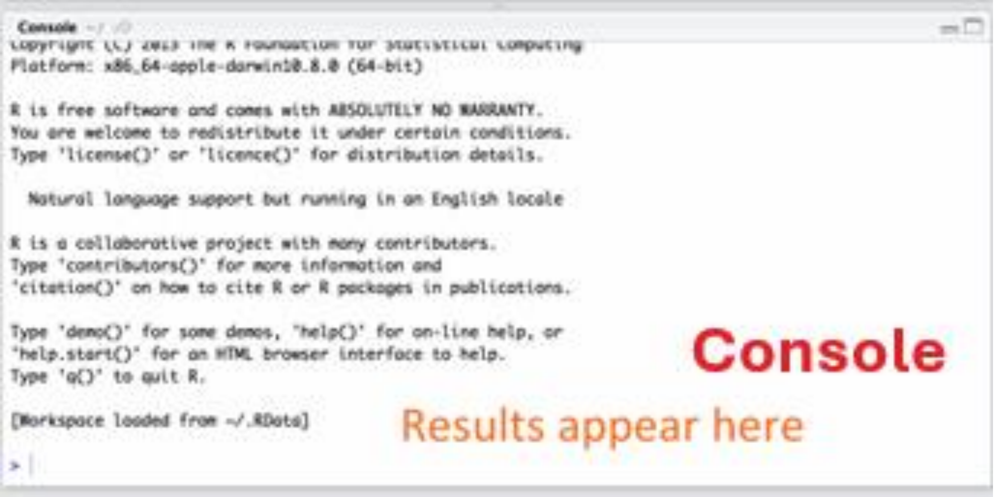
Console

Results appear here

Files/Plots/Packages/Help

Packages, plots, files

RStudio Interface



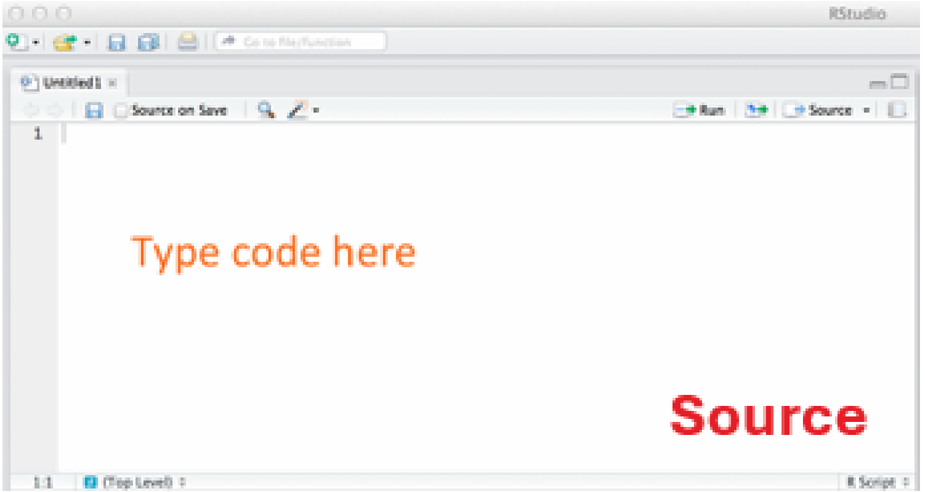
```
Copyright (C) 2013 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin10.8.0 (64-bit)

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.

[Workspace loaded from ~/.RData]

>
```

Console

Results appear here

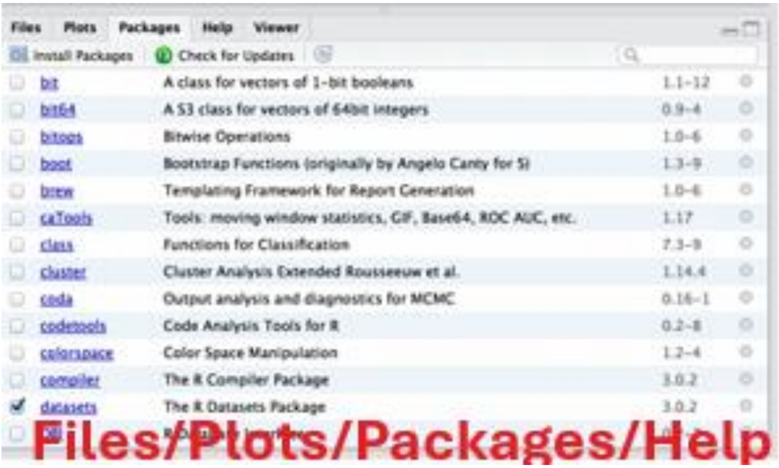
- ‘Console’ is where you will find the output of your coding and computations
- Try `10 + 5` and hit Enter/Return
- `Ctrl + L` for clear the console of all text

RStudio Interface



- ‘Source’ can be understood as any type of file, e.g. data, programming code, notes, etc.
  - Write script
  - Open programming code, e.g. an R script
  - Open other text-based file formats, e.g. .txt, .md, .tex, .bib
  - Edit scripts with code in it
  - Run the analysis you have written

RStudio Interface

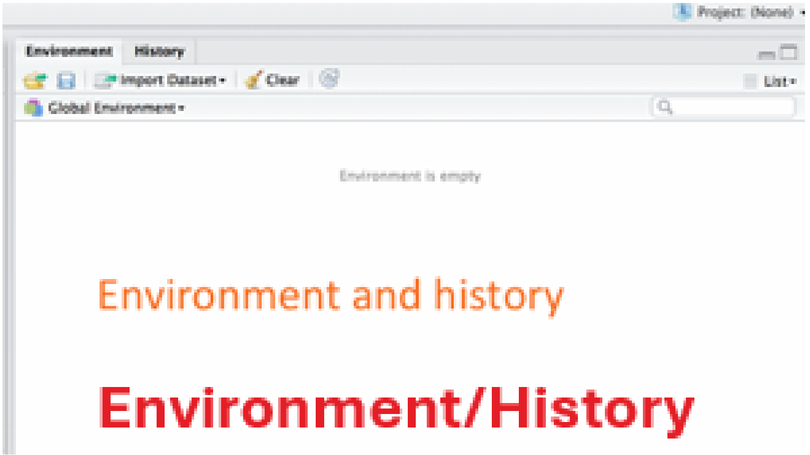


□	bit	A class for vectors of 1-bit booleans	1.1-12
□	bit64	A S3 class for vectors of 64bit integers	0.9-4
□	bitops	Bitwise Operations	1.0-6
□	boot	Bootstrap Functions (originally by Angelo Canty for S)	1.3-9
□	brew	Templating Framework for Report Generation	1.0-6
□	caTools	Tools: moving window statistics, GIF, Base64, ROC AUC, etc.	1.17
□	class	Functions for Classification	7.3-9
□	cluster	Cluster Analysis Extended Rousseeuw et al.	1.14.4
□	coda	Output analysis and diagnostics for MCMC	0.16-1
□	codetools	Code Analysis Tools for R	0.2-8
□	colorspace	Color Space Manipulation	1.2-4
□	compiler	The R Compiler Package	3.0.2
☑	datasets	The R Datasets Package	3.0.2

Files/Plots/Packages/Help

- ‘Files’ lists all the files and folders in your root directory
- ‘Plots’ is designed to show you any plots you have created using R
  - Type `boxplot(mtcars$hp)` ; call a function `boxplot()` on a dataset named mtcars and a variable hp
- ‘Packages’ are additional tools you can import and use
  - Like app on your phone
- ‘Help’; or `?mtcars` in the console
  - Google/stackov
- ‘Viewer’ – more than 2d showcase

RStudio Interface



Environment and history

Environment/History

- ‘Environment’ and shows you objects which are available for computation
- ‘History’ store whatever computation you run in the console
  - `10+5` example, and return/To console/To source
- ‘Connections’ allows you to tap into external databases directly (Not use)
- ‘Tutorial’ to find additional materials to learn R and RStudio

R Language

Guess what this does

Z <- read.csv(“MyFile.csv”)

object = function on(arguments)

An object can be many different things:

- a dataset,
- the results of a computation,
- a plot,
- a series of numbers,
- a list of names,
- a function,
- etc.

R Language

Data types and data structures

(data is stored in R as a vector)

Data types

- Numeric
- Character
- Logical

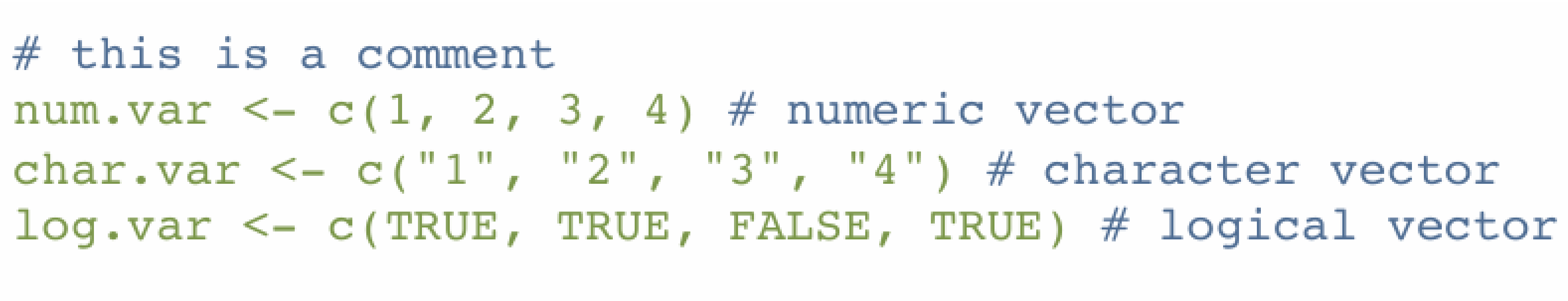
Data structures

- Vectors
- Lists
- Matrices
- Data frames

R Language

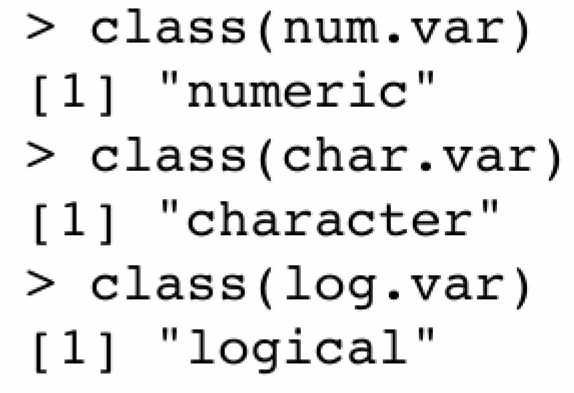
Vector

Values can be combined into vectors using the c() function.



```
# this is a comment
num.var <- c(1, 2, 3, 4) # numeric vector
char.var <- c("1", "2", "3", "4") # character vector
log.var <- c(TRUE, TRUE, FALSE, TRUE) # logical vector
```

Vectors have a *class* which determines how functions treat them

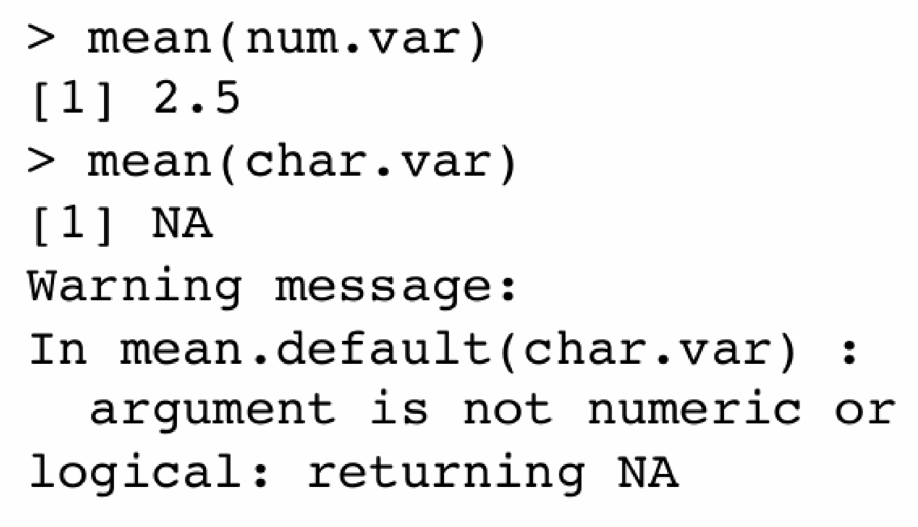


```
> class(num.var)
[1] "numeric"
> class(char.var)
[1] "character"
> class(log.var)
[1] "logical"
```

R Language

Vectors and data classes

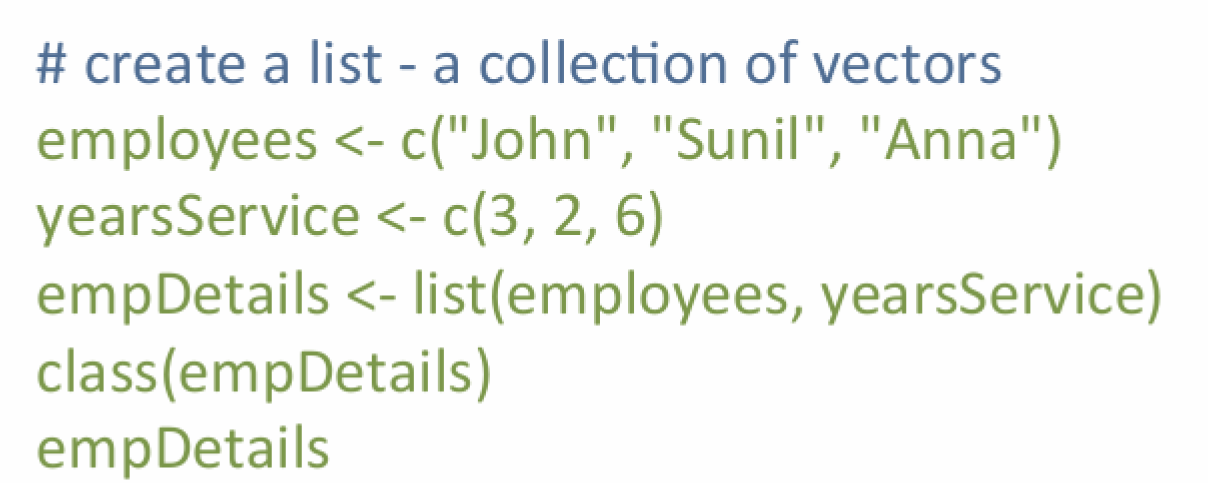
Can calculate mean of a numeric vector, but not of a character vector.



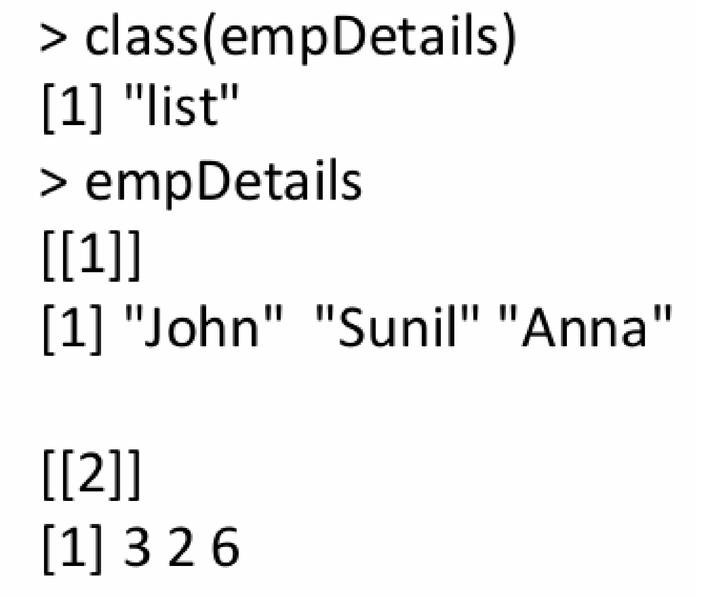
```
> mean(num.var)
[1] 2.5
> mean(char.var)
[1] NA
Warning message:
In mean.default(char.var) :
  argument is not numeric or
logical: returning NA
```

R Language

Lists



```
# create a list - a collection of vectors
employees <- c("John", "Sunil", "Anna")
yearsService <- c(3, 2, 6)
empDetails <- list(employees, yearsService)
class(empDetails)
empDetails
```

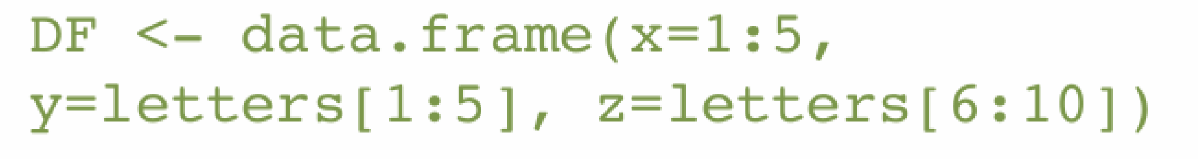


```
> class(empDetails)
[1] "list"
> empDetails
[[1]]
[1] "John"  "Sunil" "Anna"

[[2]]
[1] 3 2 6
```

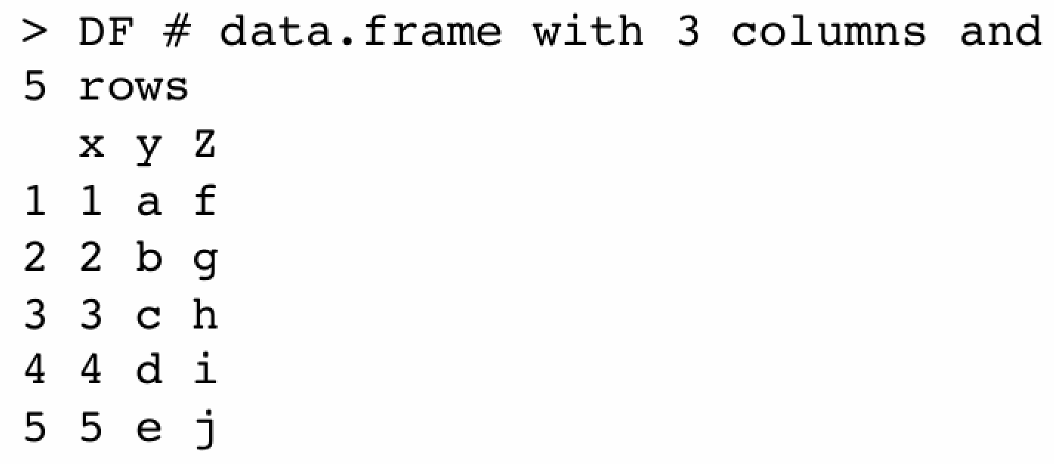
R Language

Dataframes

A data.frame is a list of vectors, each of the same length



```
DF <- data.frame(x=1:5,
y=letters[1:5], z=letters[6:10])
```

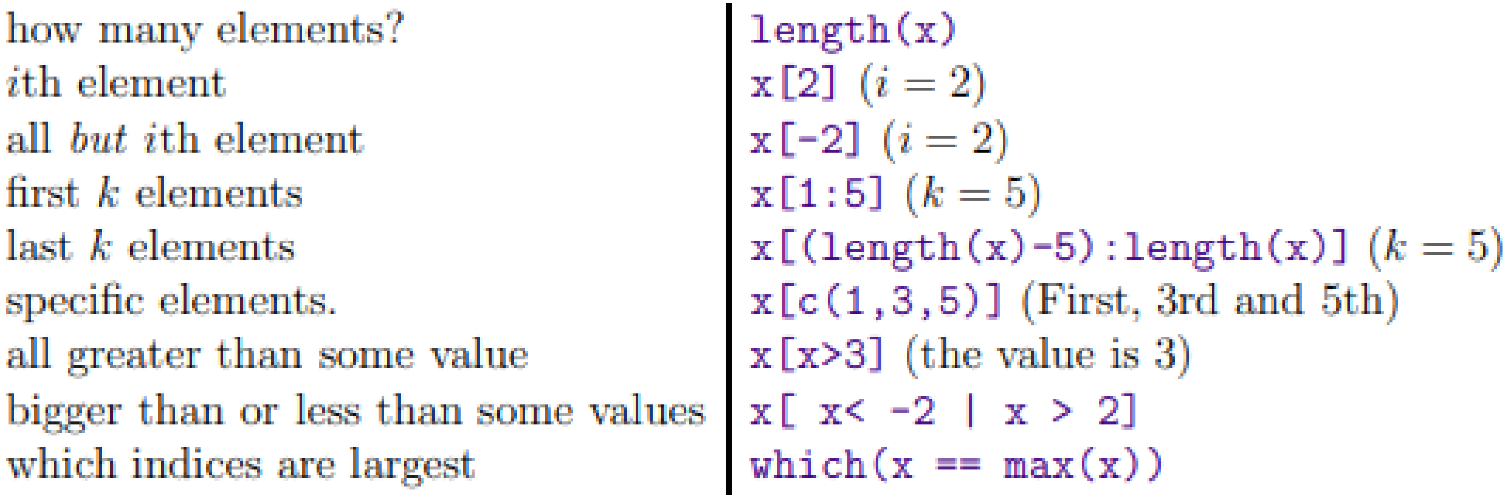


```
> DF # data.frame with 3 columns and
5 rows
  x y Z
1 1 a f
2 2 b g
3 3 c h
4 4 d i
5 5 e j
```

R Language

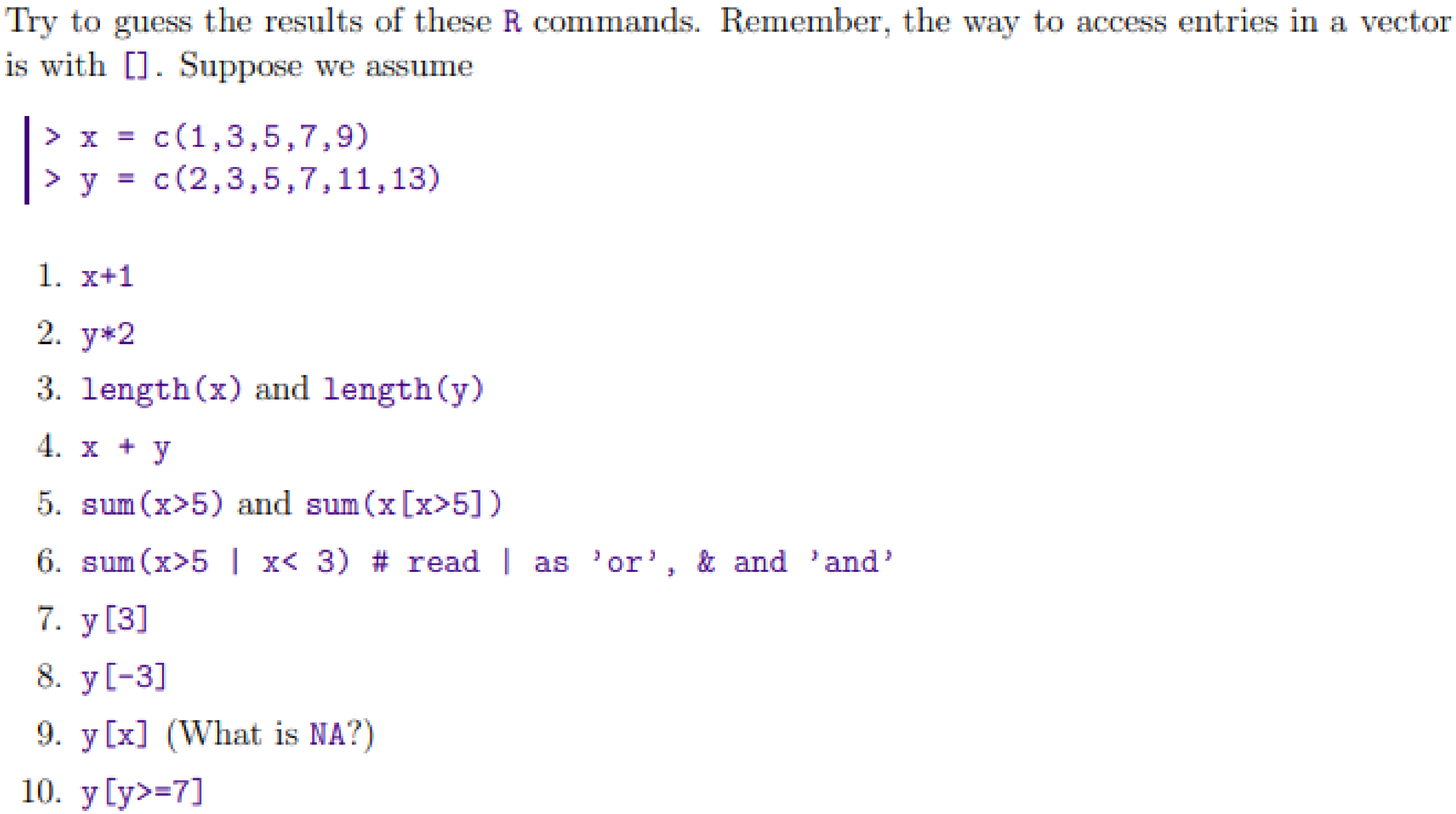
Assessing Data

There are several ways to extract data from a vector. Suppose *x* is the data vector; for example, *x=1:10*.



how many elements?	`length(x)`
ith element	`x[2]` ($i = 2$)
all *but* ith element	`x[-2]` ($i = 2$)
first k elements	`x[1:5]` ($k = 5$)
last k elements	`x[(length(x)-5):length(x)]` ($k = 5$)
specific elements.	`x[c(1,3,5)]` (First, 3rd and 5th)
all greater than some value	`x[x>3]` (the value is 3)
bigger than or less than some values	`x[ x< -2 | x > 2]`
which indices are largest	`which(x == max(x))`

Practice time



Try to guess the results of these R commands. Remember, the way to access entries in a vector is with []. Suppose we assume

```
> x = c(1,3,5,7,9)
> y = c(2,3,5,7,11,13)
```

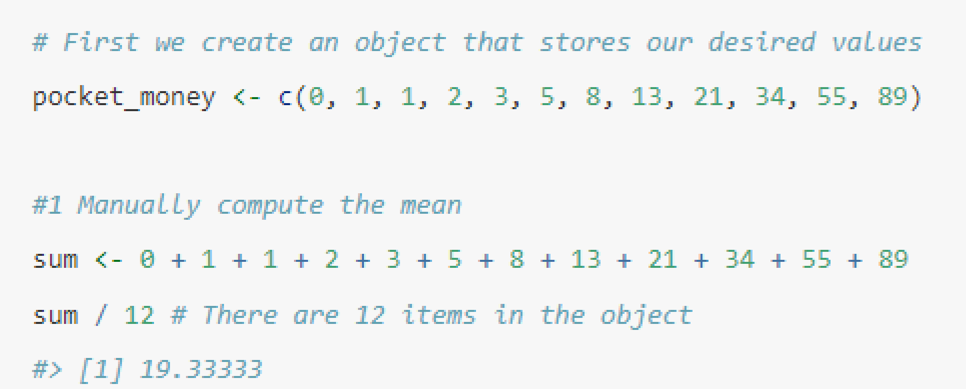
1. x+1
2. y*2
3. length(x) and length(y)
4. x + y
5. sum(x>5) and sum(x[x>5])
6. sum(x>5 | x< 3) # read | as 'or', & and 'and'
7. y[3]
8. y[-3]
9. y[x] (What is NA?)
10. y[y>=7]

R Function

Function

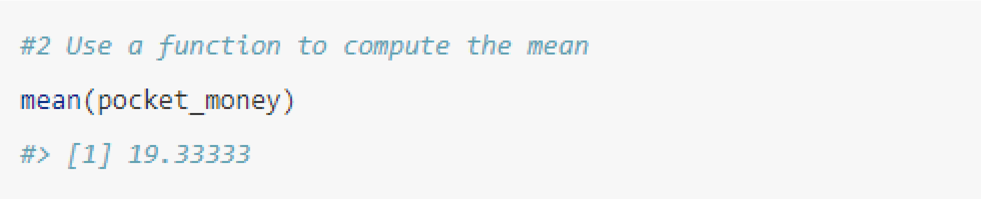
- contain lines of code that someone has written, or we have written ourselves
- shortcuts for our programming
- increase the speed with which we perform our analysis

Manually



```
# First we create an object that stores our desired values
pocket_money <- c(0, 1, 1, 2, 3, 5, 8, 13, 21, 34, 55, 89)

#1 Manually compute the mean
sum <- 0 + 1 + 1 + 2 + 3 + 5 + 8 + 13 + 21 + 34 + 55 + 89
sum / 12 # There are 12 items in the object
#> [1] 19.33333
```

Function



```
#2 Use a function to compute the mean
mean(pocket_money)
#> [1] 19.33333
```

R Package

R has many built-in functions that we can use right away. However, some of the most interesting ones have been developed by different programmers. To add more functions to your repertoire, you can install **R packages**.

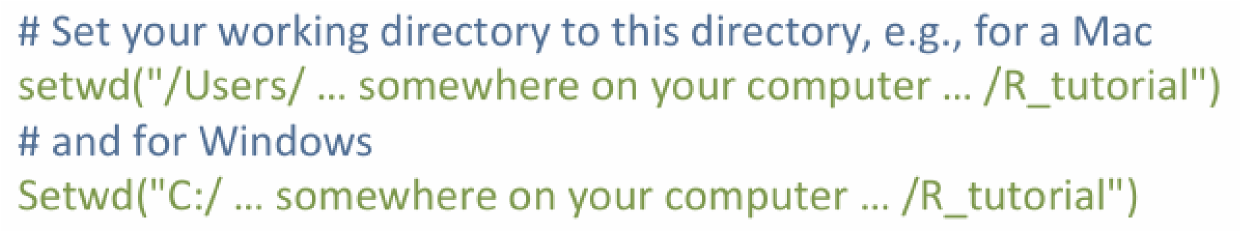
R packages are a collection of functions that you can download and use for your analysis.

- R packages do not only include functions but often include datasets and documentation of what each function does.

- What do you find those R packages?
  - Use the function *install.packages(),* or use the packages pane in RStudio.
- How to use these R packages?
  - To load an R package, we have to use the function *library().*

R Language

Set the working directory



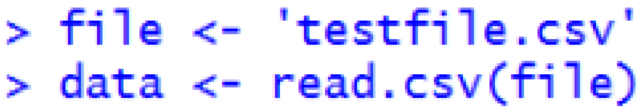
```
# Set your working directory to this directory, e.g., for a Mac
setwd("/Users/ … somewhere on your computer … /R_tutorial")
# and for Windows
Setwd("C:/ … somewhere on your computer … /R_tutorial")
```

or by click…

And check: *getwd()*

R Data

Read the data



```
> file <- 'testfile.csv'
> data <- read.csv(file)
```

- The simplest way is to export and import files in **csv** (comma separated values) format – the lingua franca of data

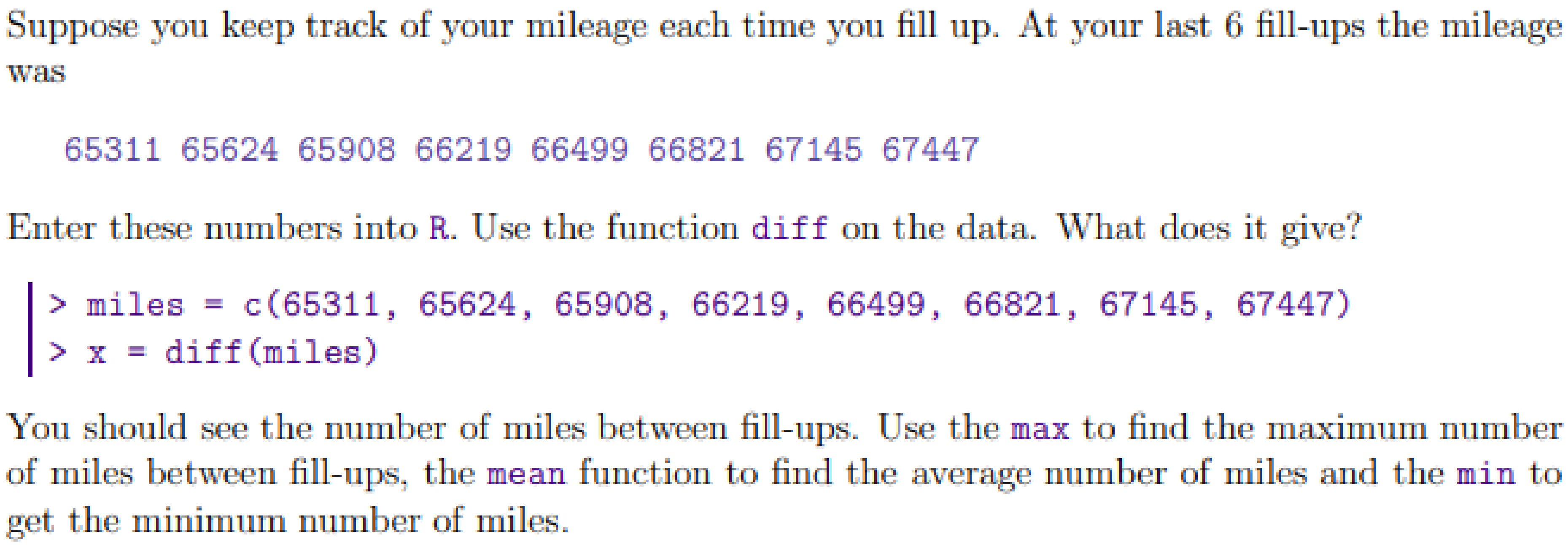
R

To give you another analogy,

- R is like a global supermarket where everyone can offer their products,
- RStudio is like my shopping cart where I can put the products I like, and
- R packages are the products I can pick from the shelves.

Luckily, R are free to use, so we do not have to bring credit card.

Practice time



Suppose you keep track of your mileage each time you fill up. At your last 6 fill-ups the mileage was

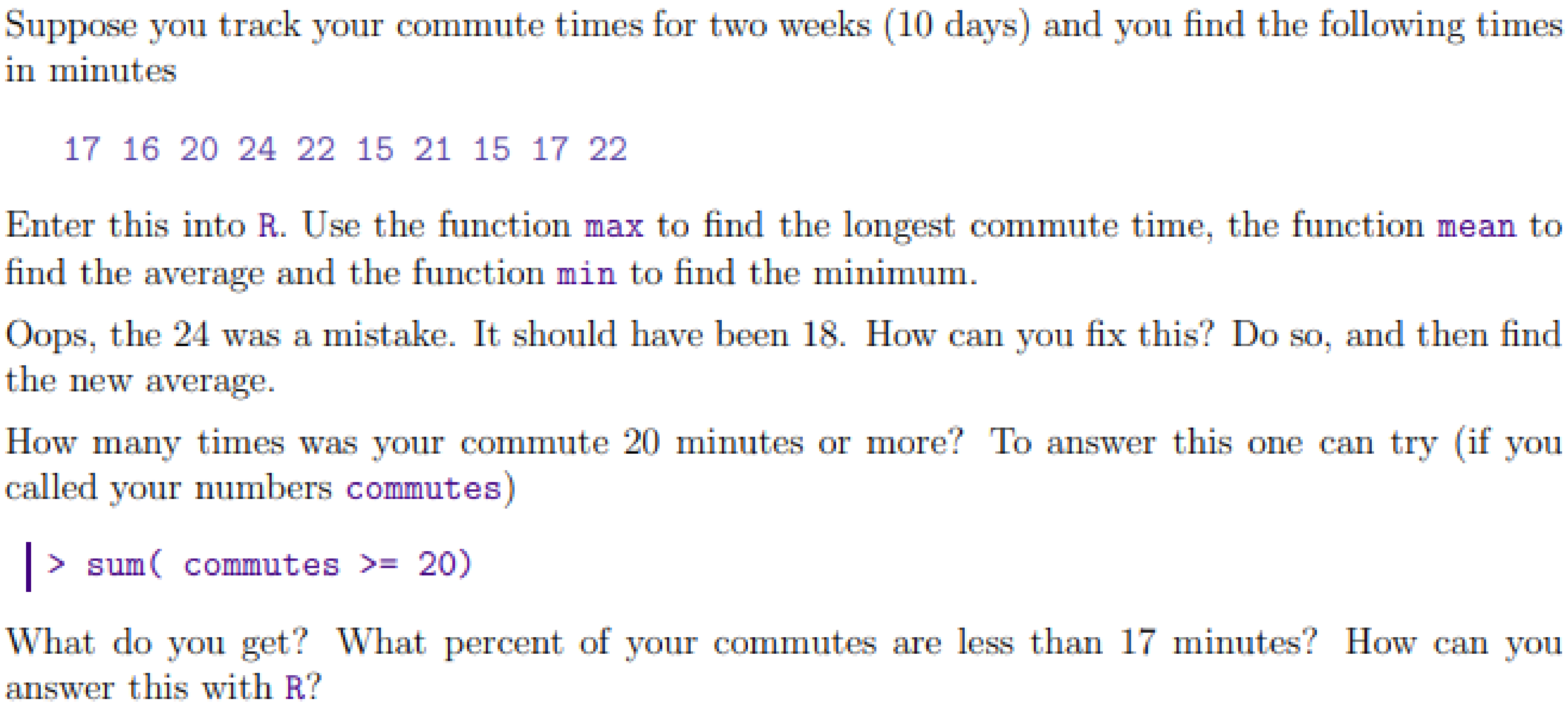
```
65311 65624 65908 66219 66499 66821 67145 67447
```

Enter these numbers into R. Use the function diff on the data. What does it give?

```
> miles = c(65311, 65624, 65908, 66219, 66499, 66821, 67145, 67447)
> x = diff(miles)
```

You should see the number of miles between fill-ups. Use the max to find the maximum number of miles between fill-ups, the mean function to find the average number of miles and the min to get the minimum number of miles.

Practice time



Suppose you track your commute times for two weeks (10 days) and you find the following times in minutes

```
17 16 20 24 22 15 21 15 17 22
```

Enter this into R. Use the function max to find the longest commute time, the function mean to find the average and the function min to find the minimum.

Oops, the 24 was a mistake. It should have been 18. How can you fix this? Do so, and then find the new average.

How many times was your commute 20 minutes or more? To answer this one can try (if you called your numbers commutes)

```
> sum( commutes >= 20)
```

What do you get? What percent of your commutes are less than 17 minutes? How can you answer this with R?

END OF LECTURE!