Introduction to Statistics

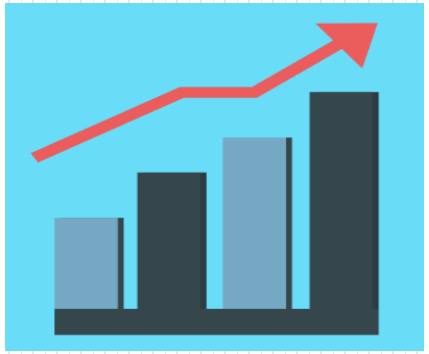
Dr. Farman Ali

Assistant Professor

DEPARTMENT OF SOFTWARE

SEJONG UNIVERSITY

Lecture-1



Introduction to Statistics Course Overview

- **Staff**
  - Farman Ali (farmankanju@sejong.ac.kr)

- **Lecture Location &Time**
  - **Lecture Class 002:Tuesday -Thursday 16:30 ~ 18:00 / Room B111, Innovation Center**
  - **Lecture Class 003: Tuesday -Thursday 18:00 ~ 19:30 / Room B111, Innovation Center**

- **Grading Policy**

  Midterm(25%), Final exam(40%), Attendance (10%), Homework + Class Quizzes +Presentation+ Assignments(25%)

- **Cheating Policy**
  - Automatic **F** for both

Introduction to Statistics Course Overview

Course Description: The statistics course is an introductory course in probability and statistics emphasizing applications in science and engineering. This course deals with various statistical tools and ideas to collect, analyze, and draw inference from data arising from both observational and experimental studies in science and engineering. The aim of the course is to give you an introduction to the concepts in probability and provide you with a basic idea of statistical inference.

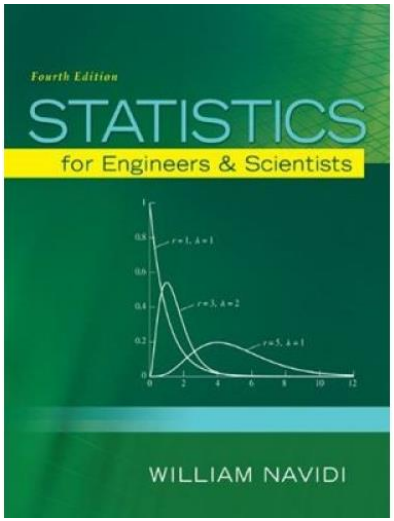
Course Notes:

- Lectures will be conducted for 3 hours per week.
- Office hours (Room: 451):
  - Monday 13:00~17:30
  - Tuesday 13:00~17:30
- Class activities and Quizzes will be used for continuous assessments.
- Notes will be uploaded to (https://ecampus.sejong.ac.kr/dashboard.php)

Introduction to Statistics Course Overview

Textbook:

Statistics for Engineers and Scientists (Fourth Edition), by William Navidi. (ISBN: 9780073401331, Publisher: McGraw-Hill Education)



Additional Guide:

- Lecture materials, assignments will be based on contents taken from recommended books and internet.
- Quizzes will be based on the material delivered during the lecture.
- The lecture material will be formatted in the form of PPT slides or hand-out notes.
- Both PPT slides and white-board will be used during lecture to discuss topics and solve equations. Slides and hand-out notes will be provided before the lecture time.

The classes will be face-to-face

Course Syllabus

- Introduction to the course
- Sampling and data presentation
- Basic of probability
- Distributions
- Confidence intervals
- Hypothesis testing
- Correlation and simple linear regression
- Multiple regression

Outline

- Introduction to statistics
- Why study statistics
- Types of statistics
- Basic concepts
- The software environment for this course

Introduction

- What is Statistics?

Statistics is the science of conducting studies to

- Collect
- Organize
- Summarize
- Analyze
- And draw conclusions from data

OR

- It is the study of the principles and the methods used in collecting, presenting, analyzing, and interpreting numerical data.

- The word statistics is derived from the Latin word Status, which is loosely defined as a statesman.

Example of Statistics

- It was reported that violent crimes were down by 3.5% in 2010 in the world
- It was reported that the average student loan debt was about $28,000.
- The college stress and mental illness poll reported that 85% of college and university students reported feeling stress daily; 75% reported stress from school work, and 64% experienced stress from grades.

Why Study Statistics?

- Data are everywhere and every day we are bombarded with different types of data and claims.
- Statistical techniques are used to make many decisions that affect our lives
- No matter what your career, you will make professional decisions that involve data. An understanding of statistical methods will help you make these decisions efectively

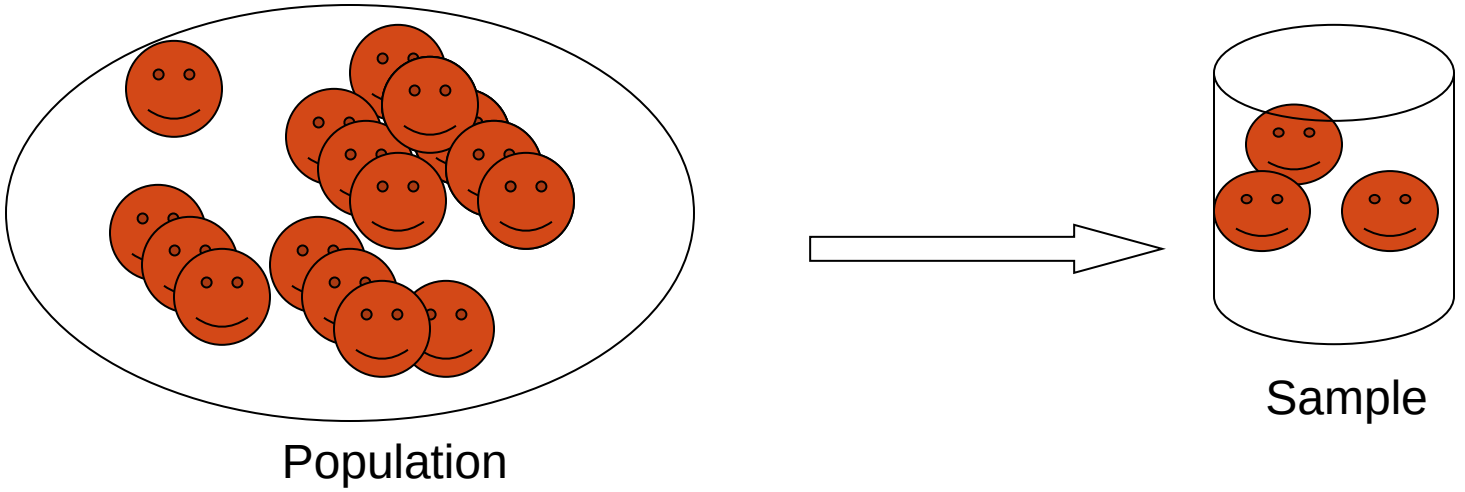
Data

Statistical techniques

decisions

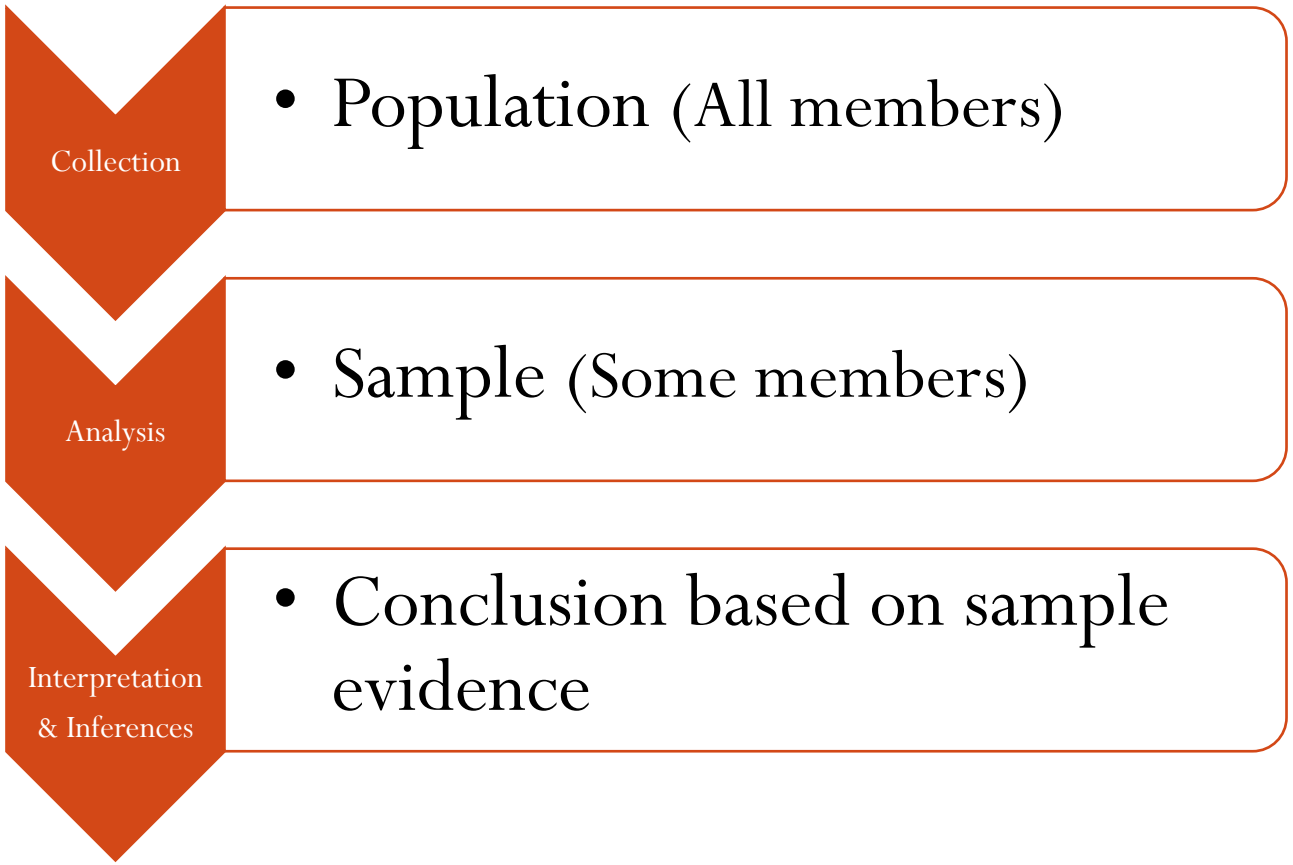
Types of Statistics

- **Statistics** is divided into two main areas, depending on how data are used.
- **Descriptive statistics** – Methods of collecting, organizing, summarizing, and presenting data in an informative way.
- **Inferential statistics** – The methods used to determine something about a population on the basis of a sample
  - **Population** – The entire set of individuals or objects of interest or the measurements obtained from all individuals or objects of interest
  - **Sample** – A portion, or part, of the population of interest



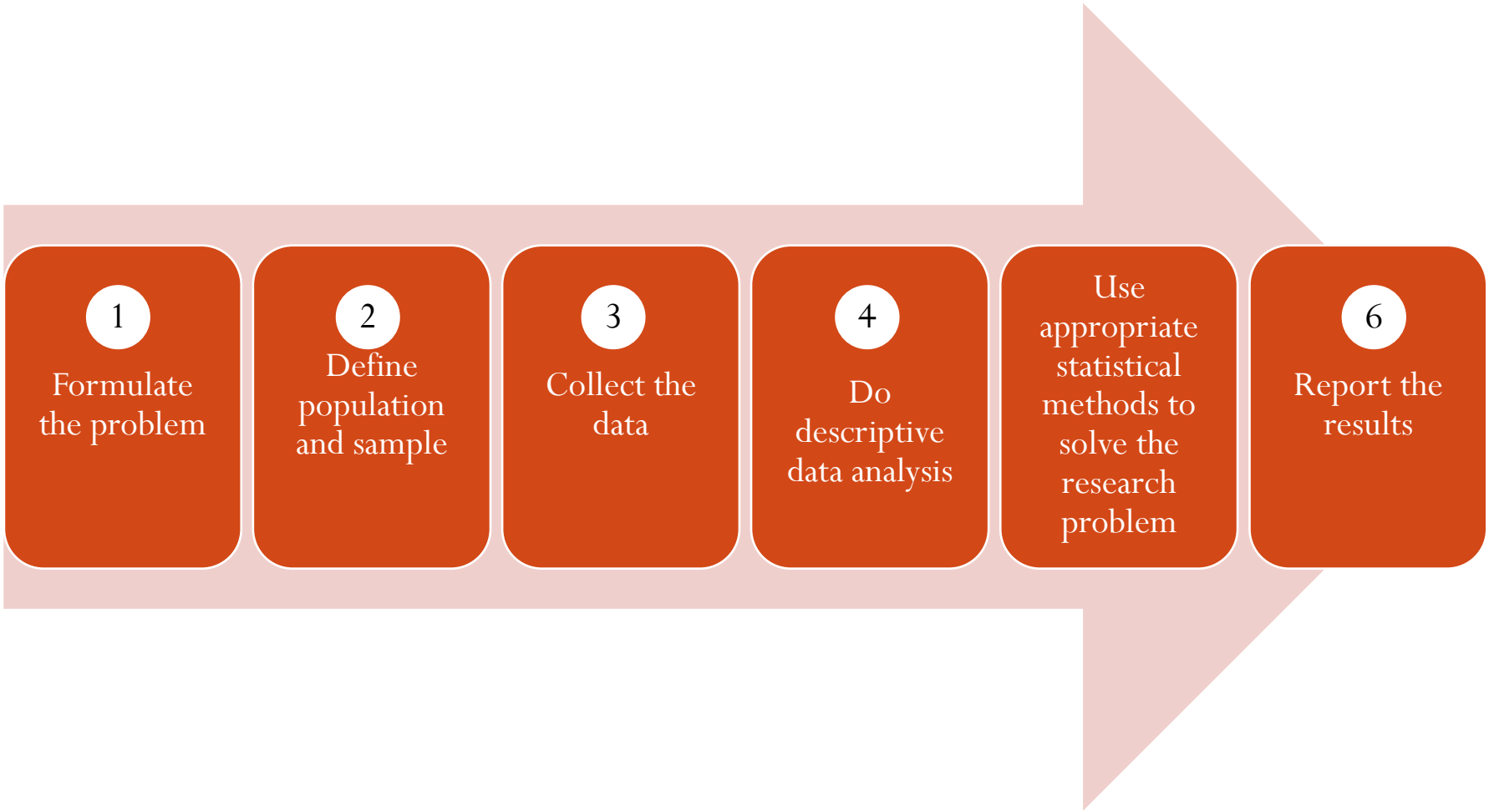
Statistics

- **Statistics** is the field of study concerned with the collection, analysis, and interpretation of uncertain data.

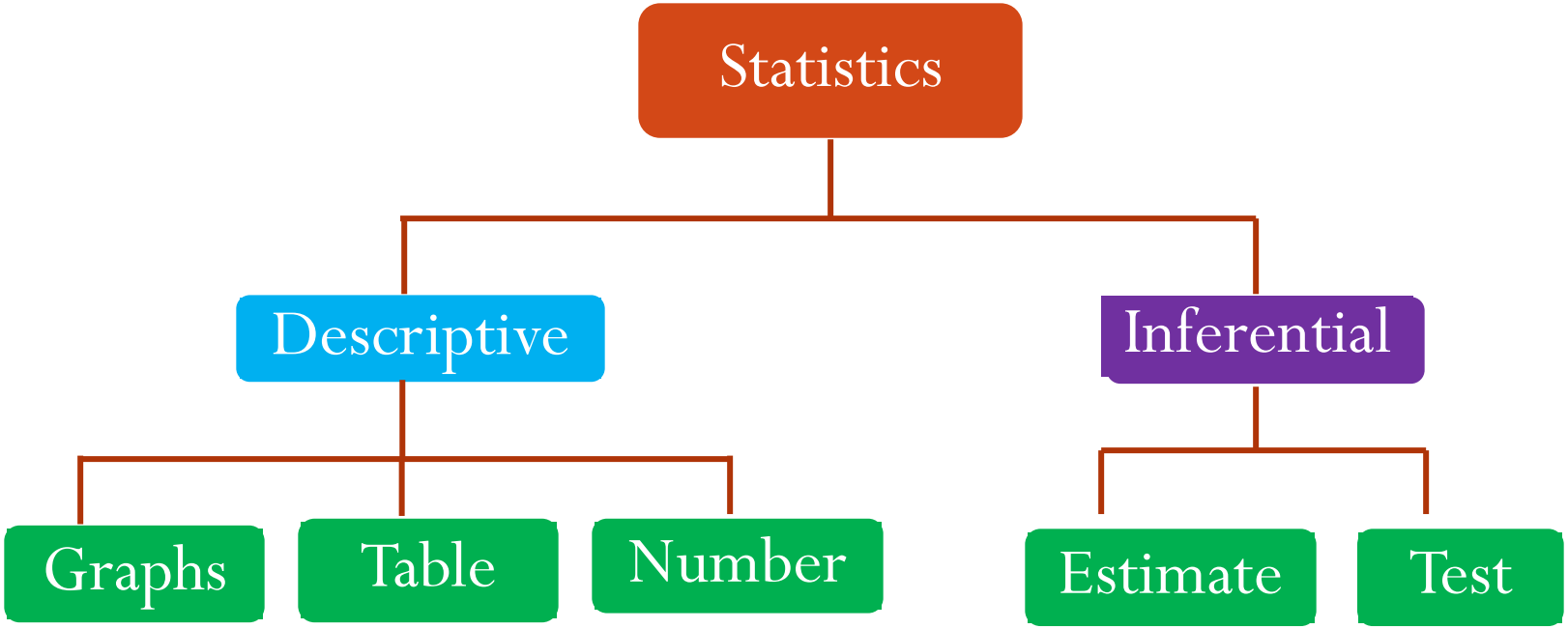


Statistics for Data Analysis

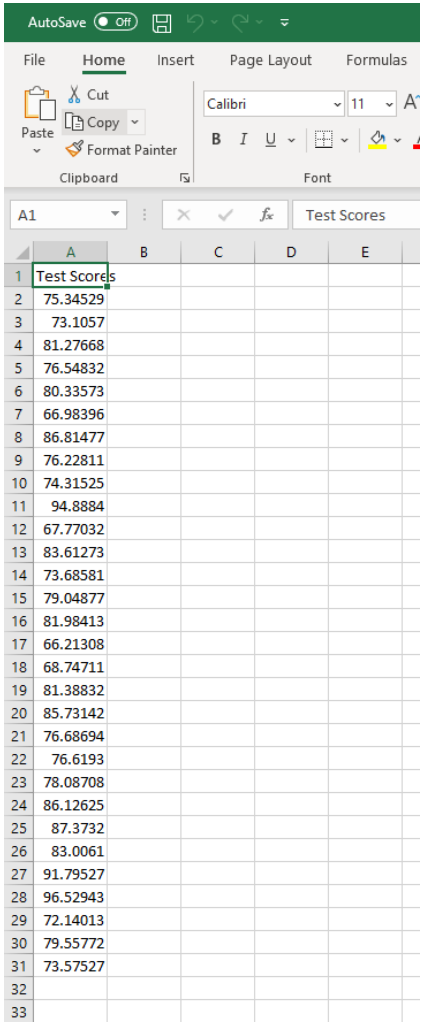
- The goal of statistics is to gain understanding from data. Any data analysis should contain following steps



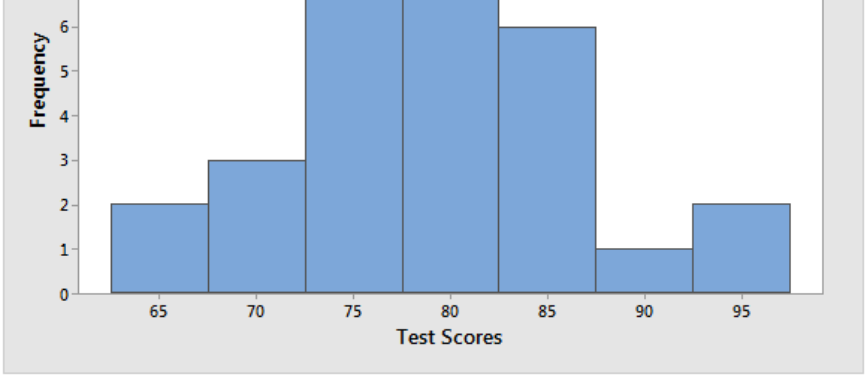
Statistics


Statistics
Descriptive
Inferential
Graphs
Table
Number
Estimate
Test

Descriptive statistics example



	A	B	C	D	E
1	Test Scores				
2	75.34529				
3	73.1057				
4	81.27668				
5	76.54832				
6	80.33573				
7	66.98396				
8	86.81477				
9	76.22811				
10	74.31525				
11	94.8884				
12	67.77032				
13	83.61273				
14	73.68581				
15	79.04877				
16	81.98413				
17	66.21308				
18	68.74711				
19	81.38832				
20	85.73142				
21	76.68694				
22	76.6193				
23	78.08708				
24	86.12625				
25	87.3732				
26	83.0061				
27	91.79527				
28	96.52943				
29	72.14013				
30	79.55772				
31	73.57527				
32					
33					



Statistic	Class value
Mean	79.18
Range	66.21 – 96.53
Proportion >= 70	86.7%

Descriptive statistics example

	nominal	metric	metric	nominal	metric	nominal	ordinal
Cases	**Gender**	**Salary**	**Age**	**Place**	**Weight**	**Company**	**Academic degree**
1	Female	1,500	33	Chicago	80	BMW	Bachelor
2	Female	1,200	33	Chicago	82.5	Ford	No
3	Male	2,200	34	New York	100.8	BMW	Bachelor
4	Male	2,100	42	New York	90	BMW	Master
5	Female	1,500	29	Chicago	67	Ford	Master
6	Female	1,700	19	Washington	60	Ford	Master
7	Male	3,000	50	Washington	77	Ford	No
8	Male	3,000	55	Washington	77	Ford	Bachelor
9	Female	2,800	31	New York	87	Ford	Bachelor
10	Male	2,900	46	New York	70	GM	Master
11	Female	2,780	36	Washington	57	BMW	No
12	Male	2,550	48	New York	64	GM	Master
13							
14							

https://datatab.net/statistics-calculator/descriptive-statistics

The various **sub-areas of descriptive statistics** can be summarized as follows:

1) Location parameter

- Mean
- Median
- Modal value
- Sum

2) Dispersion parameter

- Standard deviation
- Variance
- Range

3) Frequency tables

4) Graphics

Descriptive statistics example

- A random sample of 10 male basketball players will be drawn, whose height will be measured in meters.
- The following table with descriptive statistics on the height of basketball players. The table shows the relevant dispersion measures and location measures.

Data

Player	Body height
1	1.62
2	1.72
3	1.55
4	1.7
5	1.78
6	1.65
7	1.64
8	1.64
9	1.66
10	1.74

Statistics

	Body height
Mean value	1.67
Median	1.655
Mode	1.64
Sum	16.7
Standard deviation	0.066
Variance	0.004
Minimum	1.55
Maximum	1.78
Range	0.23

Descriptive Statistics

- In descriptive statistics the statistician tries to describe a situation.

- Descriptive statistics give information that describes the data in some manner. For example, suppose a grocery store sells Eggs, Bread, milk and fruit. If 100 items are sold, and 30 out of the 100 were Milk, then one description of the data on the grocery store items sold would be that 30% were Milk.

- This same grocery store may conduct a study on the number of bread sold each day for one month and determine that an average of 20 bread were sold each day. The average is an example of descriptive statistics.

Descriptive Statistics cont.

- Another example, consider the national census conducted by any country in every 10 years. Results of this census give you the average age, income, gender and other feature of the population. To obtain this information, the gov must have some means to collect significant data. Once the data are collected, they organize and summarize them. Finally, they present the data in some meaningful form, such as charts, reports, graph and table, etc.

- A graphical representation of data is another method of descriptive statistics. Examples of this visual representation are histograms, bar graphs and pie graphs etc. Using these methods, the data is described by compiling it into a graph, table or other visual representation.

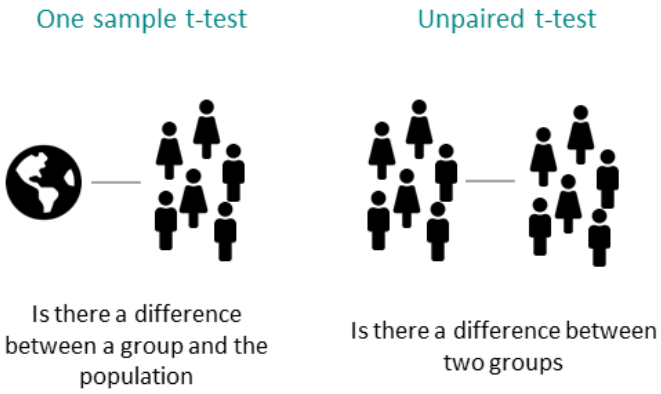
Inferential Statistics

- **Inferential statistics** makes inferences about populations using data drawn form the population. Instead of using the entire population to gather the data, the statistician will collect samples form the millions of residents and make inferences about the entire population using sample.
- In inferential statistics, the answers are never 100 % accurate because the calculations use a sample taken from the population. This sample doesn't include every measurement from the population.

Simple test procedures	Regression Analysis	Correlation analysis	ANOVA
• t-Test • Binominal Test • Chi-square test • Mann-Whitney U Test • Wilcoxon-Test • ...	• Simple linear regression • Multiple regression • Logistic regression • ...	• Pearson Correlation analysis • Spearman Rank Correlation • ...	• Single factorial ANOVA • Two factorial ANOVA • ANOVA with measurement repetitions • ...

Inferential Statistics

- Testing of statements about the population on the basis of sample characteristics.
- Different statistical methods or hypothesis tests are used.
- In the example above, a sample of 10 basketball players was drawn and then exactly this sample was described, this is the task of descriptive statistics.
- If you want to make a statement about the population, you need the inferential statistics.
- For example, it could be of interest if basketball players are larger than the average male population. To test this hypothesis a t-Test is calculated, the t-test compares the sample mean with the mean of the population.



Inferential Vs Descriptive

Examples:

- Happiness significantly raises a person's pain level tolerance. (Inferential)
- 306, people died with Covid-19 in 2020 at South Korea (Descriptive)
- 30% people have A type blood. (Inferential)

Some Basic Concepts

Before going on, some basic concepts are required

- Population
- Sample
- Parameter and Statistic
- Data
- Variable
- Data Collection

Population: A population consists of all subjects (human or otherwise) that are studied, (all members of a defined group that we are studying or collecting information on for data driven decisions).

Examples

- All Students studying at Sejong university
- All the registered voters in South Korea
- All parts produced today

Types of population

Finite population (Countable Population): If it is possible to count all items of population.

Examples

- The number of vehicles crossing a bridge every day
- The number of births per years in a particular hospital

Size of finite population: Total number of individuals /population (N).

Infinite Population (un-countable population): it is not possible to count all items of a population.

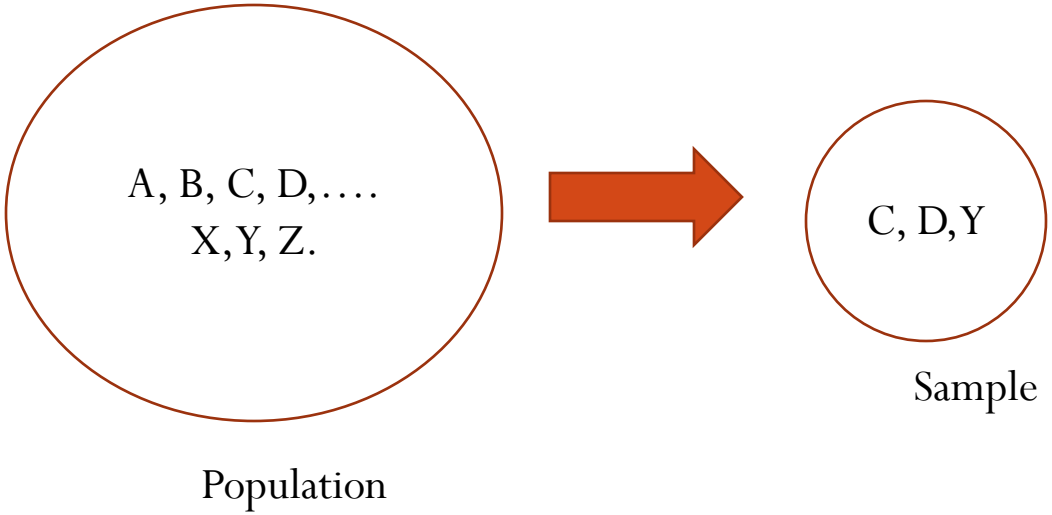
Examples

- The number of stars in the sky
- The number of germs in the body of a patient perhaps something which is uncountable.

Some Basic Concepts

Sample

A sample is a subset of the population. (A part of the population is called a sample).



Examples

- 1000 voters selected at random for interview
- A few parts selected for destructive testing
- Only software department Students are selected.

Sample size: Total number of individuals/ units in sample(n)

Some Basic Concepts

Parameter: A numerical value summarizing all the data of an entire population. e.g. Population Mean, population variance etc.

Statistic : A numerical value summarizing the sample data. e.g. Sample Mean, sample variance etc.

Example:

- Average income of all faculty members working at Sejong University is a *parameter*.
- Average income of faculty members of Software Department at Sejong University is a *statistic*.

A statistics student is interested in finding out something about the average value (in won) of cars owned by the faculty members working at Sejong University.

Question: Identify Population, Sample, parameter and statistic.

Answer:

Some Basic Concepts

Variable

- A variable is a characteristic or attribute that can assume different values.
- A characteristic that changes or varies over time for different individuals or objects under consideration.

Examples

- Hair color
- White blood cell count
- Time to failure of a computer component.

Data

- An **experimental unit** is the individual or object on which a variable is measured.
- A **measurement** results when a variable is actually measured on an experimental unit
- A set of measurements, called **data**, can be **sample** or **population**.

Some Basic Concepts

Examples 1

Variable:

- Hair color

Experimental unit:

- Person

Typical Measurements

- Brown, black, blonde, etc.

Examples 2

Variable:

- Time until a light bulb burns out

Experimental unit:

- Light bulb

Typical Measurements

- 1500 hours, etc.

How many variables we are going to measured?

- Univariate data: One variable is measured on a single experimental unit (individual or object).
- Bivariate data: Two variables are measured on a single experimental unit (individual or object).
- Multivariate data: More than two variables are measured on a single experimental unit (individual or object).

The software environment for this course

What is R?

- R is a language and environment for statistical computing and graphics.
- Designed by Ross Ihaka and Robert Gentleman
- R is an open source programming language
- Website www.r-project.org

Ross Ihaka
Professor of Statistics

Robert Gentleman
Canadian statistician

R has many uses

- Work with data: subset, merge, and transform datasets with a powerful syntax
- Analysis: use existing statistical functions like regression or write your own
- Graphics: graphs can be made quickly during analysis and polished for publication quality displays

Why learn a whole language to look at data versus Excel?

1. Recreate/redo your exact analysis
2. Automate repetitive tasks
3. Access to statistical methods not available in Excel
4. Graphs are more elegant

Why R versus SAS, SPSS, or Stata?

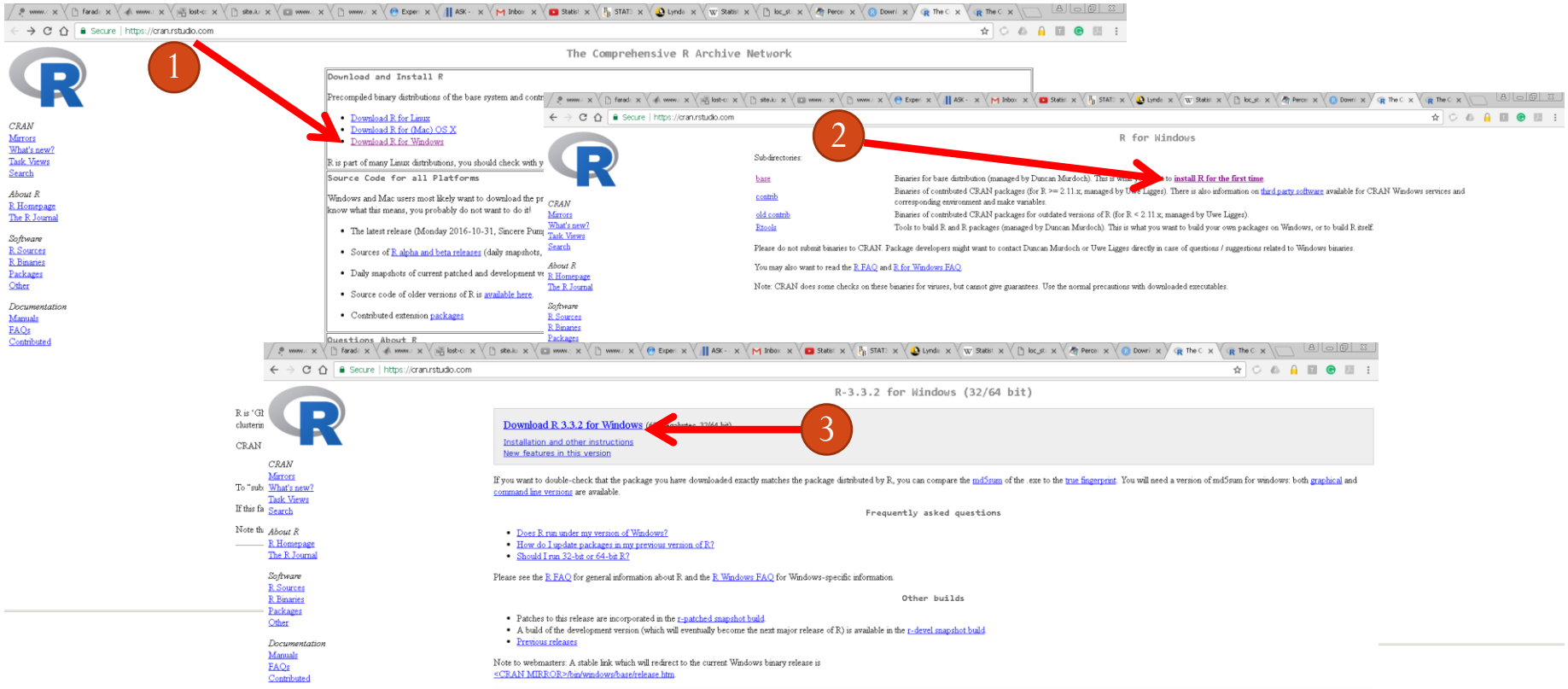
- It's free!
- It runs on Mac, Windows, and Linux
- It has state-of-the-art graphics capabilities
- It contains advanced statistical routines not yet available in other packages – a de facto standard in statistics
- Can program new statistical methods or automate data manipulation/analysis

Installing R base Package

Visit https://cran.rstudio.com/ and follow 1, 2, and 3

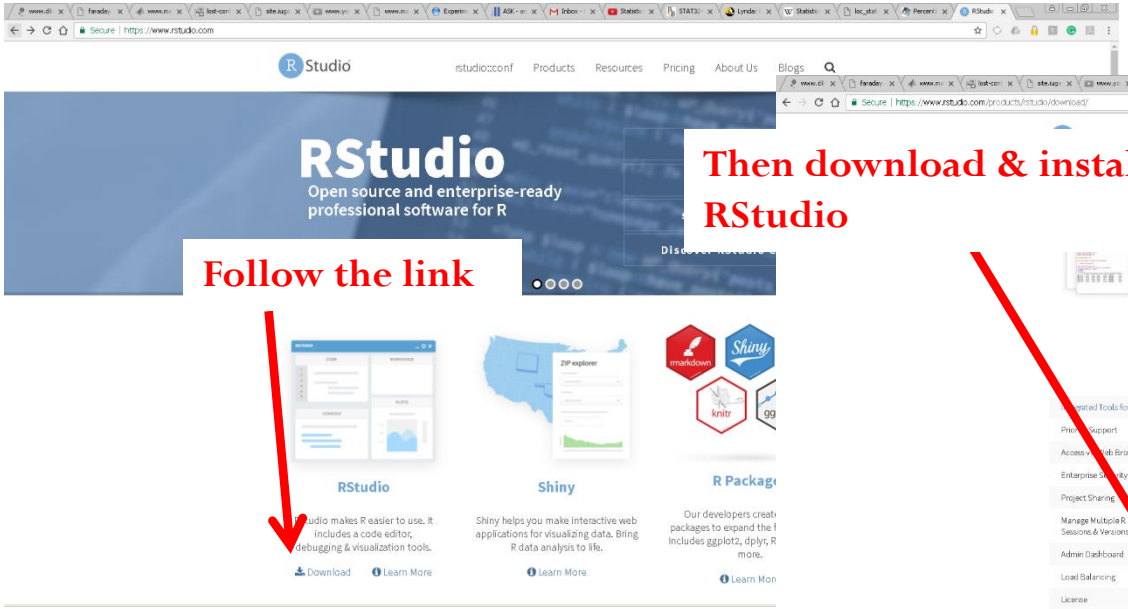
The direct link for R base package is

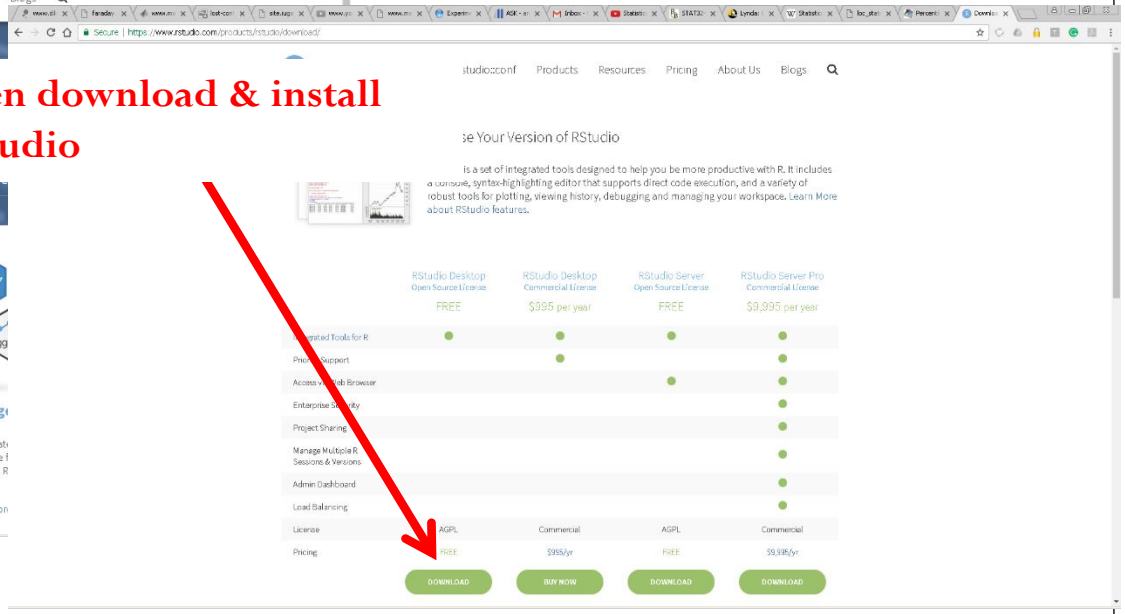
https://cran.rstudio.com/bin/windows/base/R-3.3.2-win.exe



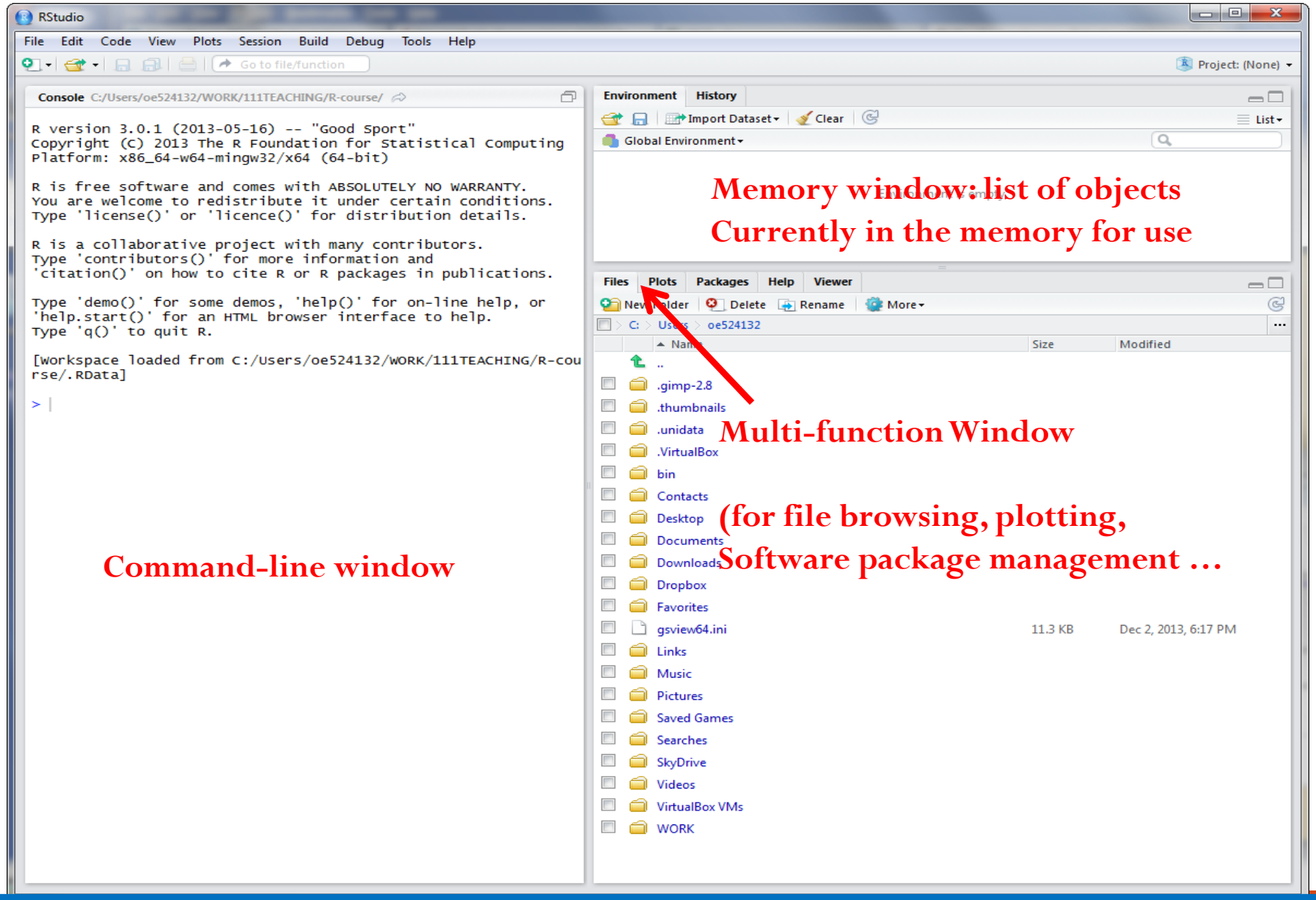
Installing R-Studio on Windows 7 (+)

- After you installed R base package, Goto http://www.rstudio.com/
- Go to Rstudio download
- Follow the link and then choose the desktop application
- Follow the instructions of the Installer

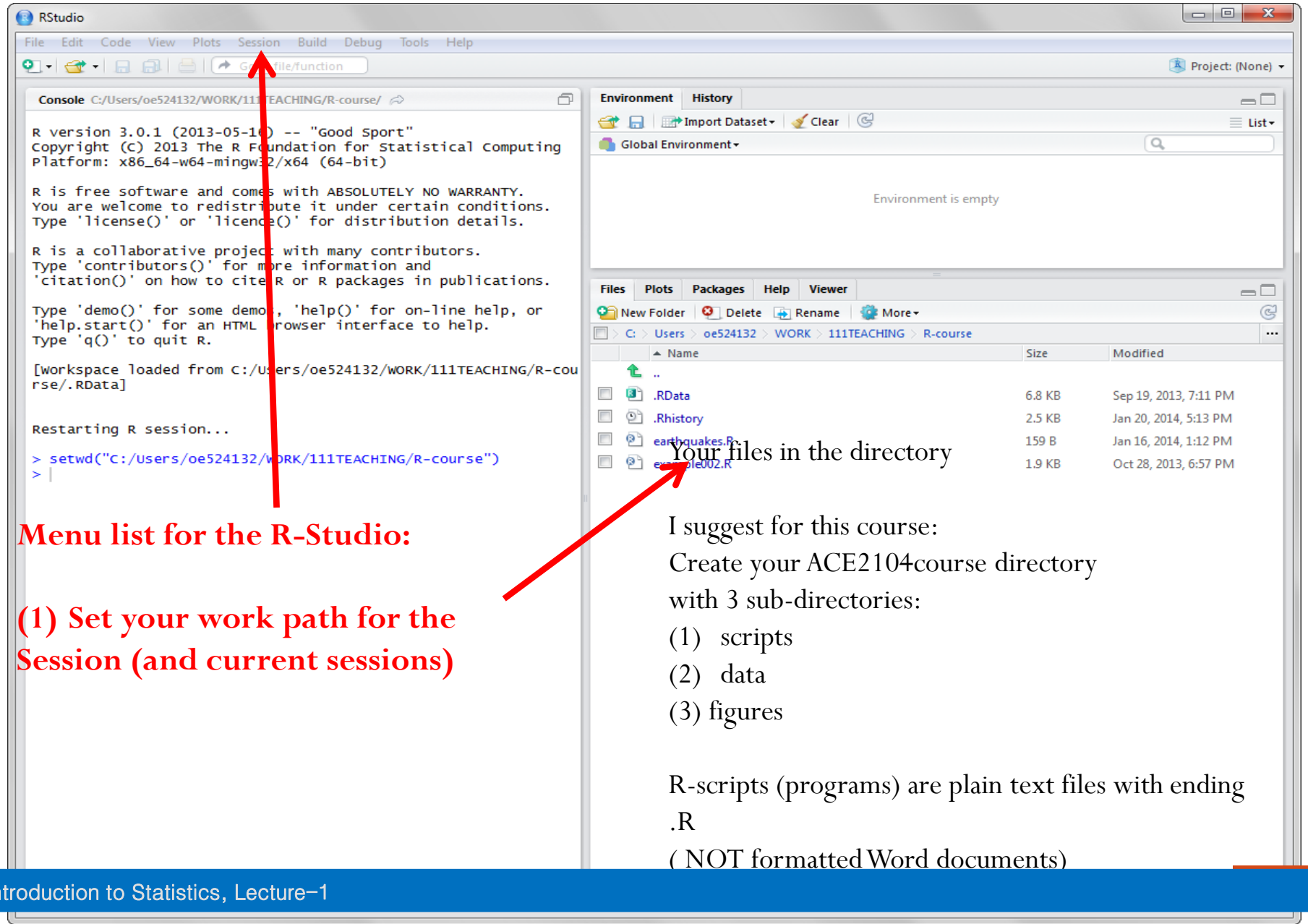




A first session with R-Studio (Windows 7)



A first session with R-Studio (Windows 7)



Menu list for the R-Studio:

(1) Set your work path for the Session (and current sessions)

Your files in the directory

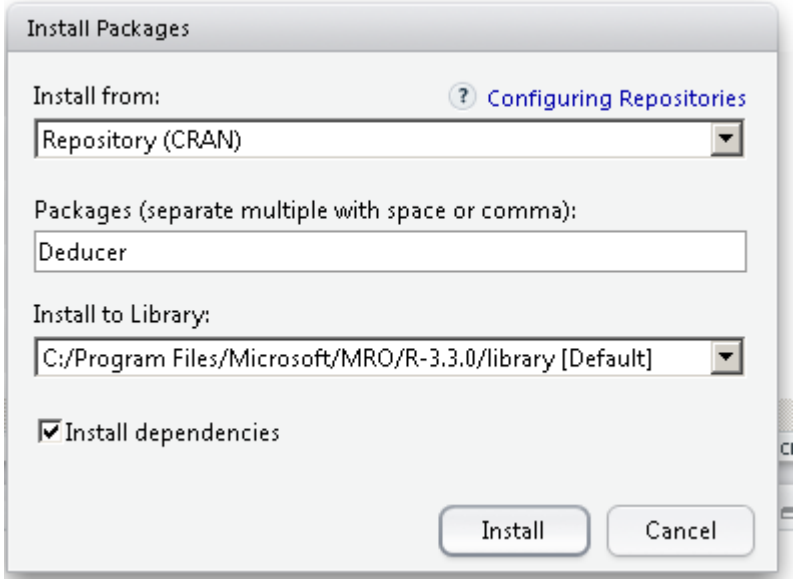
I suggest for this course:
Create your ACE2104course directory
with 3 sub-directories:

(1) scripts
(2) data
(3) figures

R-scripts (programs) are plain text files with ending .R
(NOT formatted Word documents)

Installing “Deducer” package

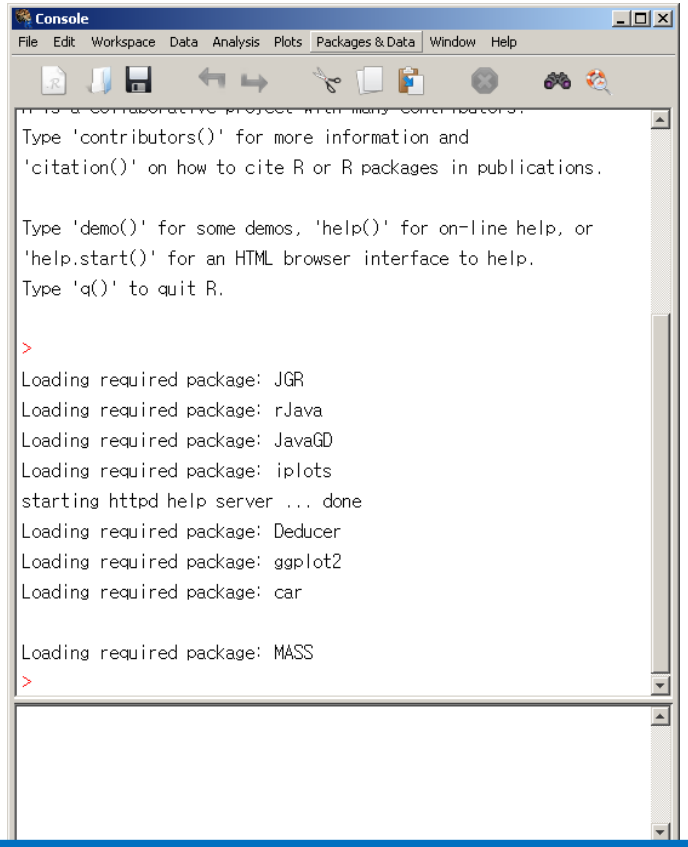
1. First install deducer from here http://www.hpmrg.org/software/Deducer-R-2.15.0-win.exe
2. Open up RStudio
3. Go to Tools> Install Packages
3. Find and select "Deducer" and choose OK.
4. This will download Deducer and the other packages which it requires, including ggplot2.
5. Then, open Tools> install Packages> write "DeducerExtras"

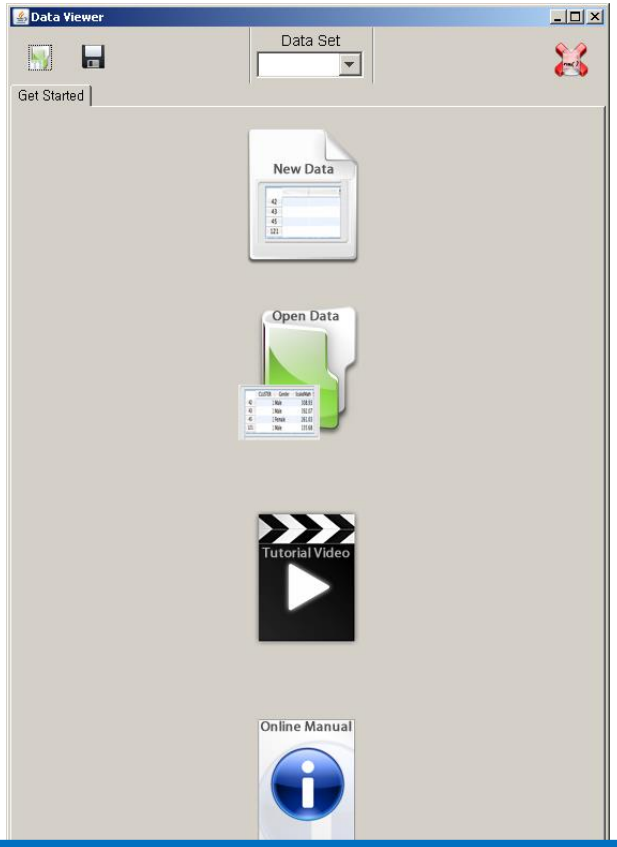


Run Duducer

- Open Rstudio
- Write at the console the following two lines:

```
library(JGR)
JGR()
```





Summary

- Statistics Course Overview
- Introduction to Statistics
- Why study statistics?
- Types of statistics
- Basic concepts
- The software environment for this course
  - Installing R base packages
  - Installing Rstudio
  - Installing Deducer


Thank
You !