# Introduction to Statistics

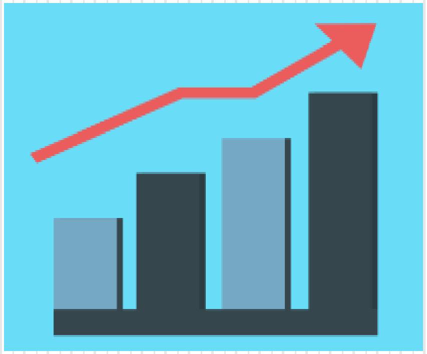
Dr. Farman Ali

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: I**nnovation Center-Room B112(Wednesday 12pm ~ 3pm)**
  - Lecture: I**nnovation Center-Room B102(Friday 3pm ~ 6pm)**

- **Grading Policy**

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

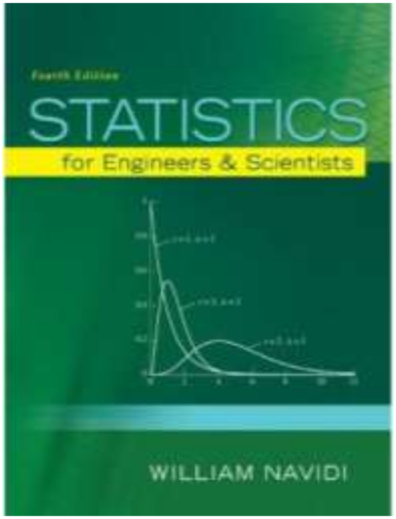
**Course Notes:**

- Lectures will be conducted for 3 hours per week.
- Office hours:
  - 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 the Sejong blackboard(blackboard.sejong.ac.kr)

# 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 class will be offline if less than 30 students registered this course. Otherwise, class will be online.

# Introduction to Statistics Course Overview

## 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

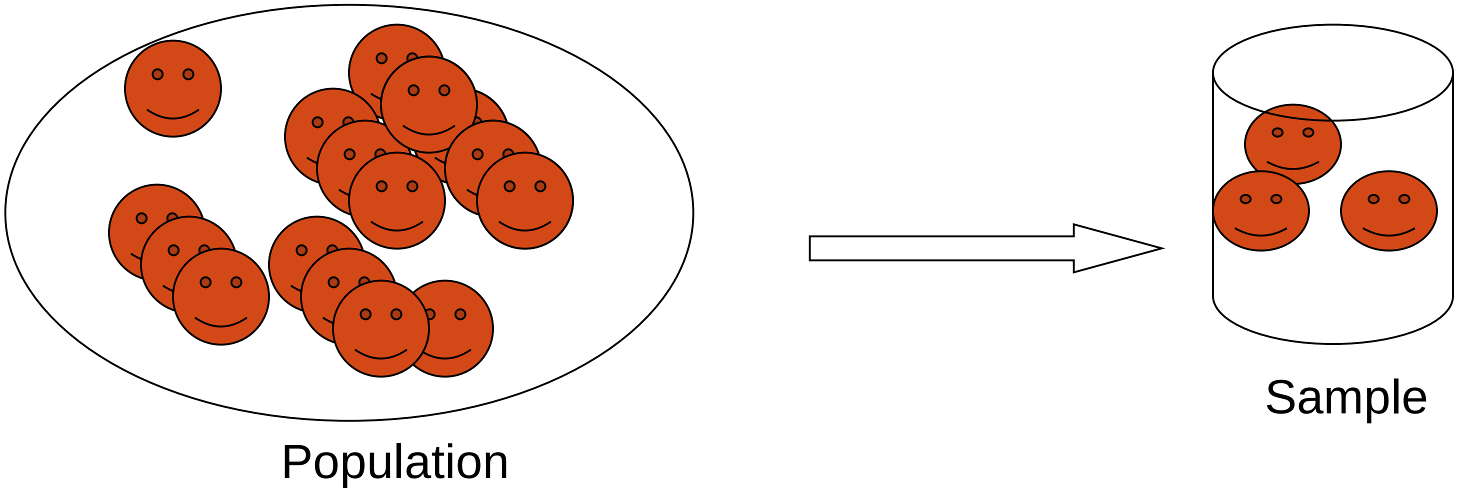
Statistical techniques

decisions

# Types of Statistics

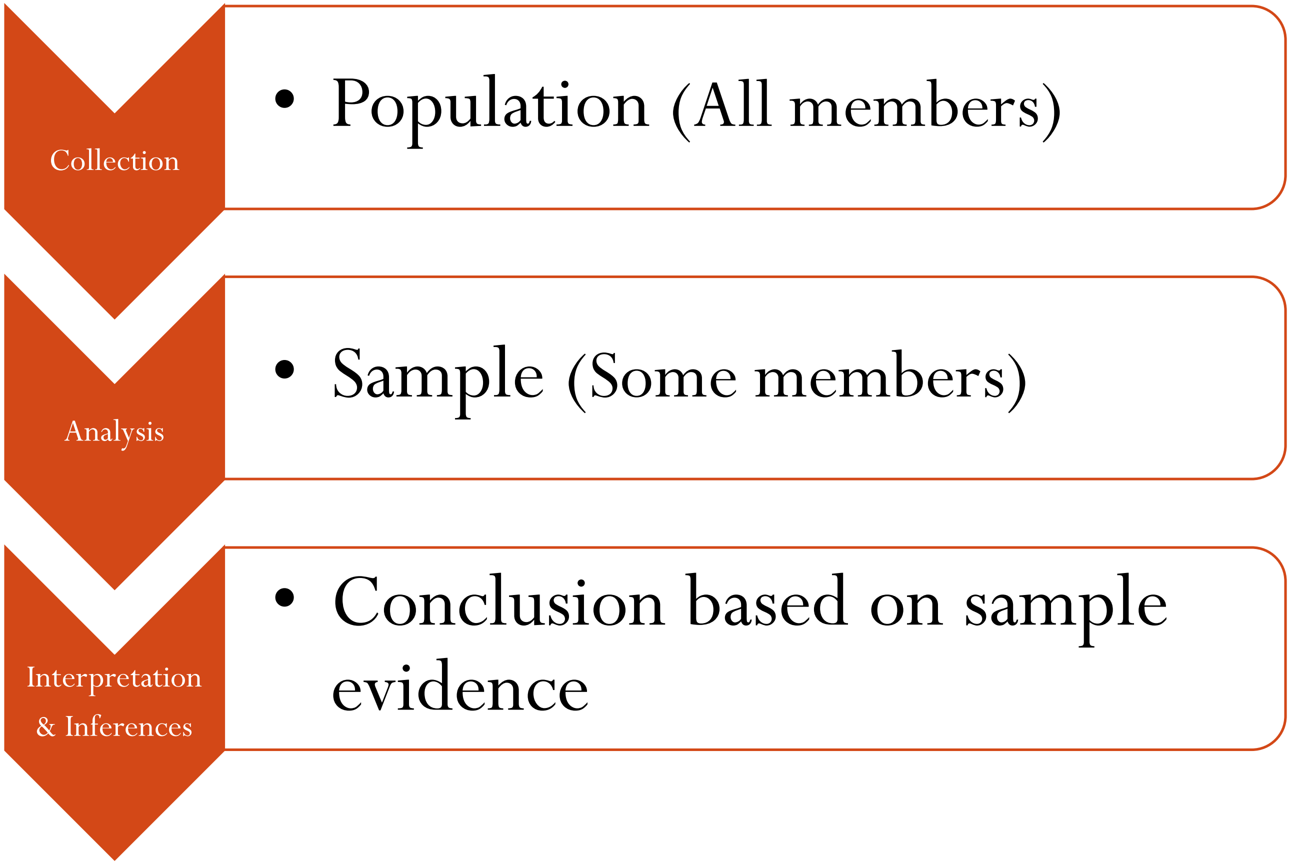
- **Statistics** is divided into two main areas, depending on how data are used.

- **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.



Collection

- Population (All members)

Analysis

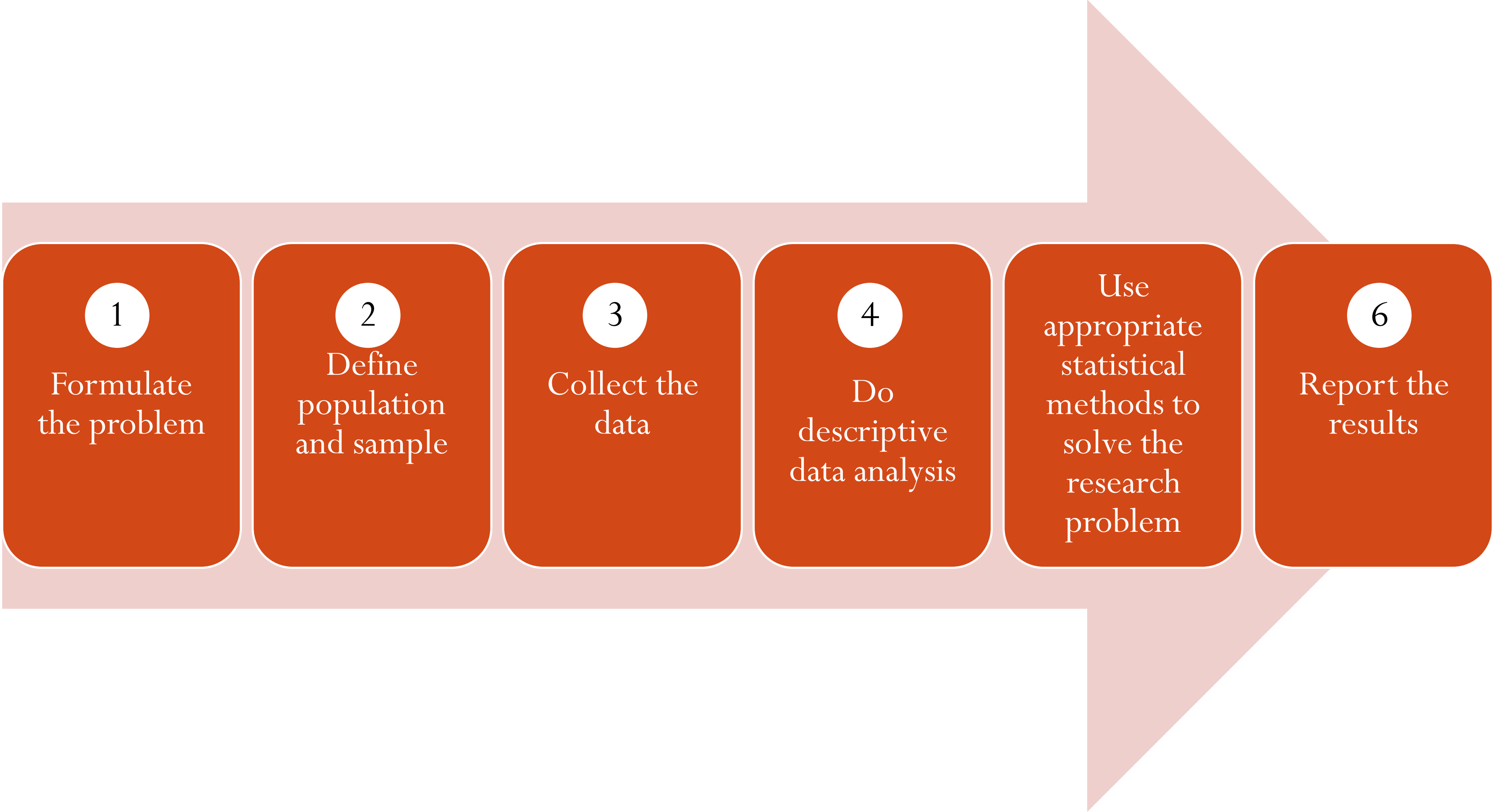
- Sample (Some members)

Interpretation & Inferences

- Conclusion based on sample evidence

# Statistics for Data Analysis

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



1 Formulate the problem

2 Define population and sample

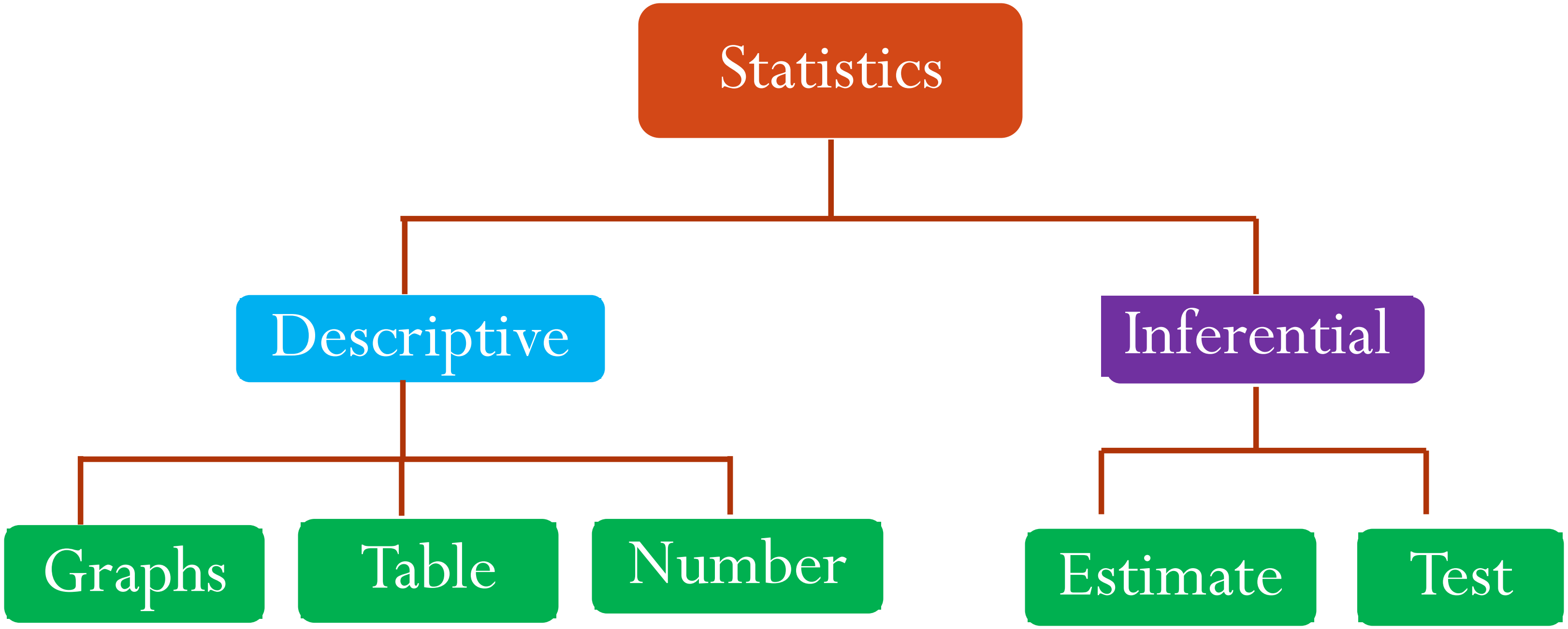
3 Collect the data

4 Do descriptive data analysis

Use appropriate statistical methods to solve the research problem

6 Report the results

# Statistics


Statistics
Descriptive
Inferential
Graphs
Table
Number
Estimate
Test

# 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.

- The Sample is a set of data taken from the population to represent the population. Probability distributions, hypothesis testing, correlation testing and regression analysis all fall under the category of inferential statistics.

- 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.

# 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
- 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

# Some Basic Concepts

## 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 data
- 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.

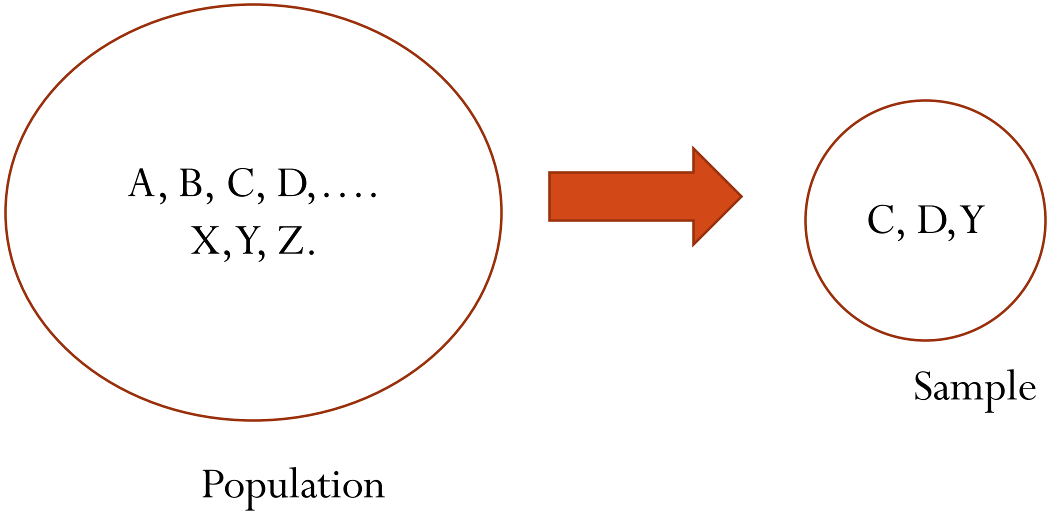
**Examples**

- 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

## 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?

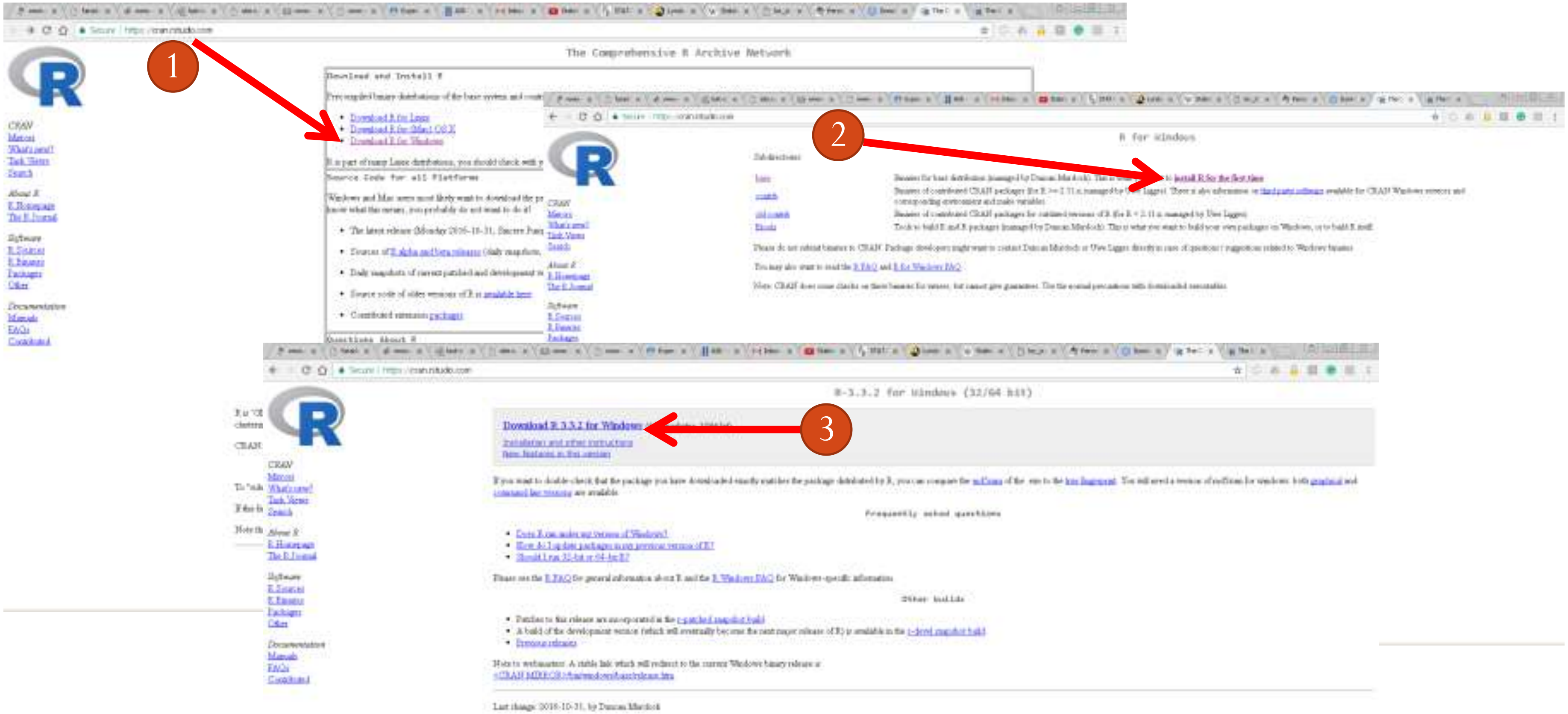
- It's free!
- It runs on Mac, Windows, and Linux
- It has state-of-the-art graphics capabilities
- It contains advanced statistical routines not yet available in other packages – a de facto standard in statistics
- Can program new statistical methods or automate data manipulation/analysis

# 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



The Comprehensive R Archive Network

Download and Install R

Precompiled binary distributions of the base system and contri

- Download R for Linux
- Download R for (Mac) OS X
- Download R for Windows

R is part of many Linux distributions, you should check with y

Source Code for all Platforms

Windows and Mac users most likely want to download the pr know what this means, you probably do not want to do it!

- The latest release (Monday 2016-10-31, Sincere Pum
- Sources of R alpha and beta releases (daily snapshots,
- Daily snapshots of current patched and development ve
- Source code of older versions of R is available here
- Contributed extension packages

Questions About R

R for Windows

Subdirectories:

| | |
|---|---|
| base | Binaries for base distribution (managed by Duncan Murdoch). This is what you want to install R for the first time. |
| contrib | Binaries of contributed CRAN packages (for R >= 2.11.x; managed by Uwe Ligges). There is also information on third party software available for CRAN Windows services and corresponding environment and make variables. |
| old contrib | Binaries of contributed CRAN packages for outdated versions of R (for R < 2.11.x; managed by Uwe Ligges). |
| Rtools | Tools to build R and R packages (managed by Duncan Murdoch). This is what you want to build your own packages on Windows, or to build R itself. |

Please do not submit binaries to CRAN. Package developers might want to contact Duncan Murdoch or Uwe Ligges directly in case of questions / suggestions related to Windows binaries.

You may also want to read the R FAQ and R for Windows FAQ.

Note: CRAN does some checks on these binaries for viruses, but cannot give guarantees. Use the normal precautions with downloaded executables.

R-3.3.2 for Windows (32/64 bit)

Download R 3.3.2 for Windows (62 megabytes, 32/64 bit)

Installation and other instructions

New features in this version

If you want to double-check that the package you have downloaded exactly matches the package distributed by R, you can compare the md5sum of the .exe to the true fingerprint. You will need a version of md5sum for windows: both graphical and command line versions are available.

Frequently asked questions

- Does R run under my version of Windows?
- How do I update packages in my previous version of R?
- Should I run 32-bit or 64-bit R?

Please see the R FAQ for general information about R and the R Windows FAQ for Windows-specific information.

Other builds

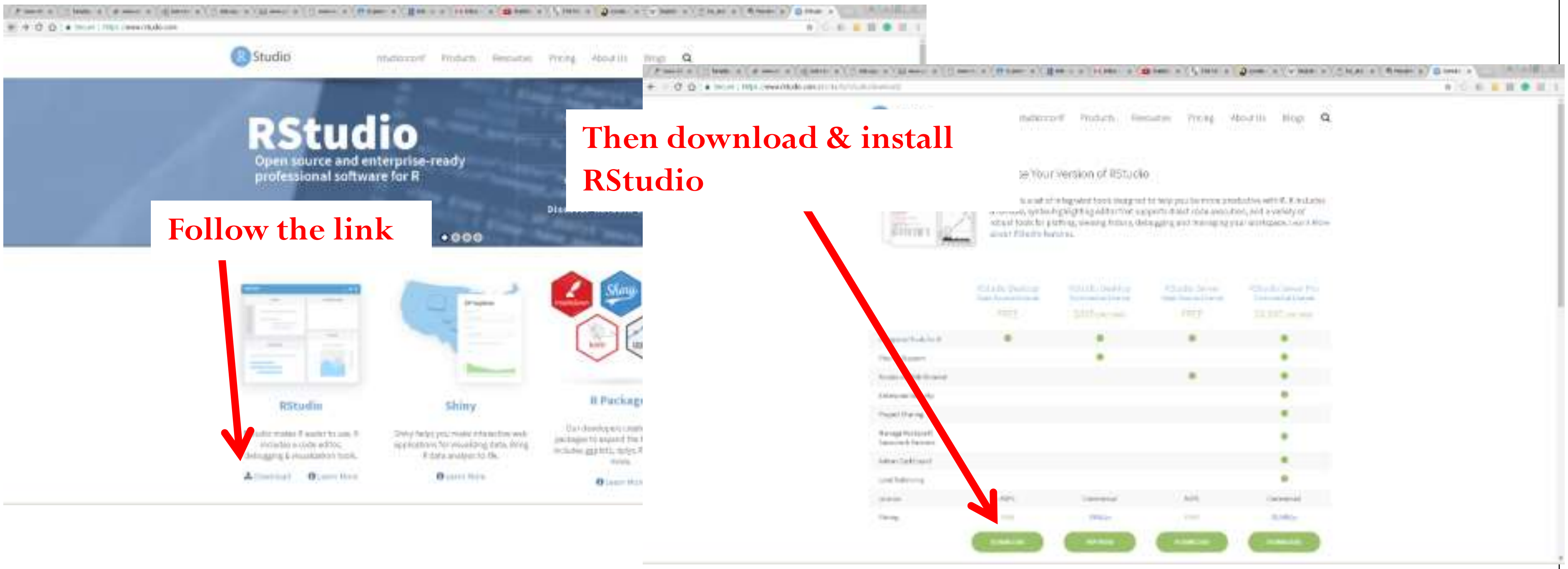
- Patches to this release are incorporated in the r-patched snapshot build.
- A build of the development version (which will eventually become the next major release of R) is available in the r-devel snapshot build.
- Previous releases

Note to webmasters: A stable link which will redirect to the current Windows binary release is
<CRAN MIRROR>/bin/windows/base/release.htm.

Last change: 2016-10-31, by Duncan Murdoch

# 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



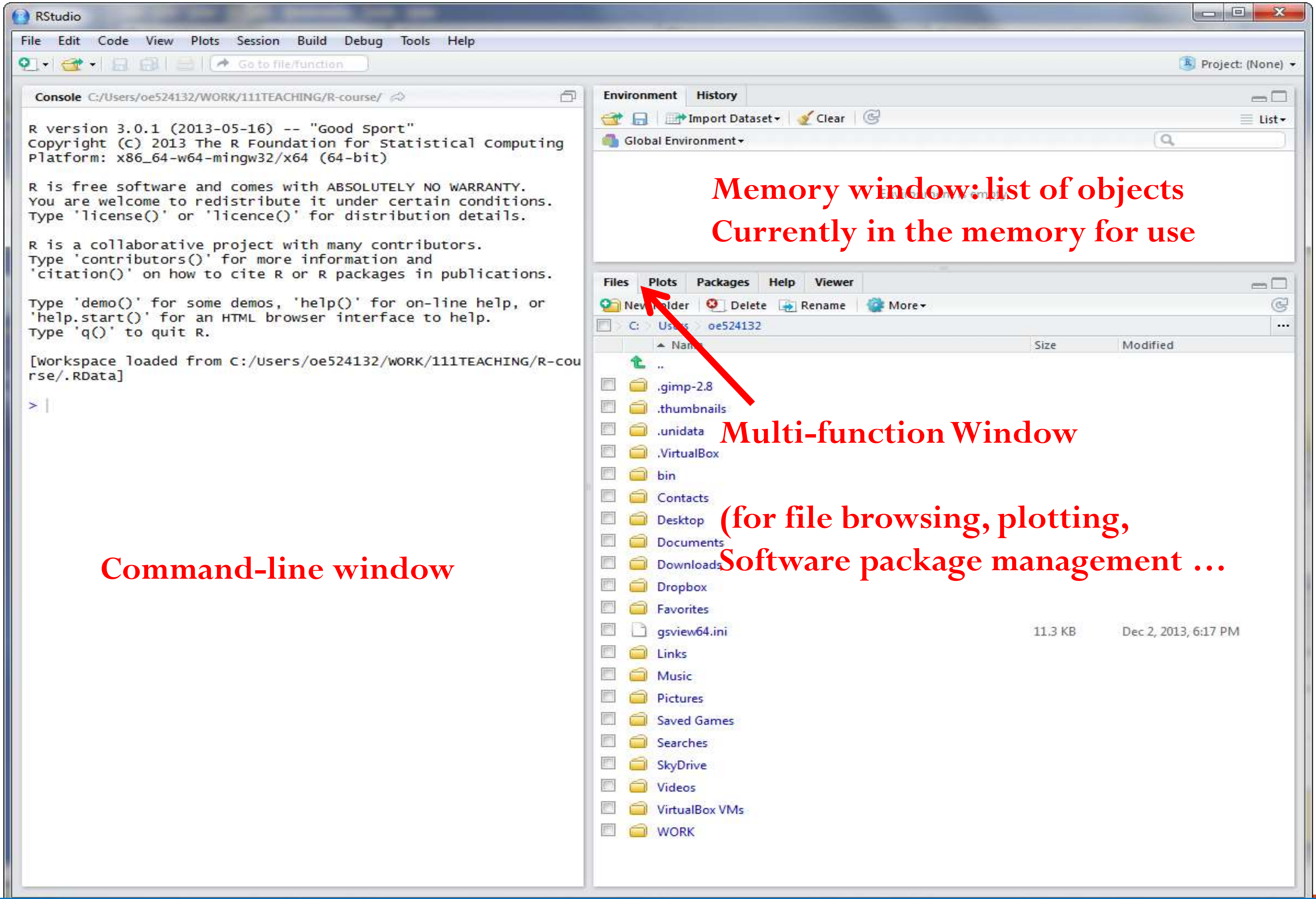
RStudio

Open source and enterprise-ready professional software for R

Follow the link

Then download & install RStudio

# A first session with R-Studio (Windows 7)



RStudio
File Edit Code View Plots Session Build Debug Tools Help
Go to file/function
Project: (None)
Console C:/Users/oe524132/WORK/111TEACHING/R-course/
Environment History
Import Dataset Clear
List
Global Environment
Files Plots Packages Help Viewer
New Folder Delete Rename More
C: Users oe524132

```
R version 3.0.1 (2013-05-16) -- "Good Sport"
Copyright (C) 2013 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.

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 C:/Users/oe524132/WORK/111TEACHING/R-cou
rse/.RData]

> |
```

| | Name | Size | Modified |
|---|---|---|---|
| | .. | | |
| □ | .gimp-2.8 | | |
| □ | .thumbnails | | |
| □ | .unidata | | |
| □ | .VirtualBox | | |
| □ | bin | | |
| □ | Contacts | | |
| □ | Desktop | | |
| □ | Documents | | |
| □ | Downloads | | |
| □ | Dropbox | | |
| □ | Favorites | | |
| □ | gsview64.ini | 11.3 KB | Dec 2, 2013, 6:17 PM |
| □ | Links | | |
| □ | Music | | |
| □ | Pictures | | |
| □ | Saved Games | | |
| □ | Searches | | |
| □ | SkyDrive | | |
| □ | Videos | | |
| □ | VirtualBox VMs | | |
| □ | WORK | | |

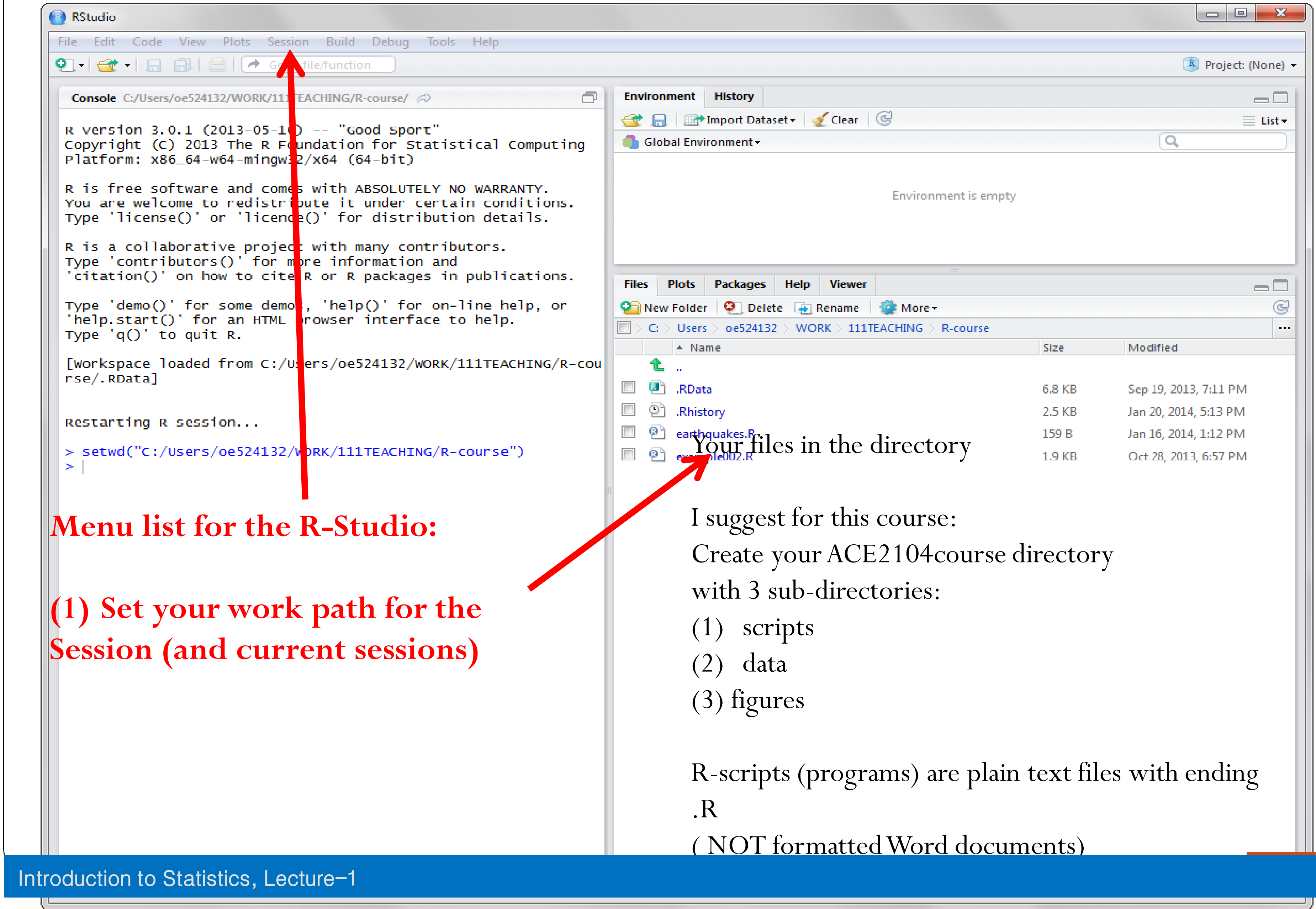
Memory window: list of objects
Currently in the memory for use

Multi-function Window

(for file browsing, plotting,
Software package management …

Command-line window

# A first session with R-Studio (Windows 7)


RStudio
File Edit Code View Plots Session Build Debug Tools Help
Go to file/function
Project: (None)
Console C:/Users/oe524132/WORK/111TEACHING/R-course/
Environment History
Import Dataset Clear
List
Global Environment
Environment is empty
Files Plots Packages Help Viewer
New Folder Delete Rename More
C: Users oe524132 WORK 111TEACHING R-course

```
R version 3.0.1 (2013-05-16) -- "Good Sport"
Copyright (C) 2013 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.

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 C:/Users/oe524132/WORK/111TEACHING/R-cou
rse/.RData]


Restarting R session...

> setwd("C:/Users/oe524132/WORK/111TEACHING/R-course")
> 
```

| Name | Size | Modified |
|---|---|---|
| .. | | |
| .RData | 6.8 KB | Sep 19, 2013, 7:11 PM |
| .Rhistory | 2.5 KB | Jan 20, 2014, 5:13 PM |
| earthquakes.R | 159 B | Jan 16, 2014, 1:12 PM |
| example002.R | 1.9 KB | Oct 28, 2013, 6:57 PM |

Your files in the directory

**Menu list for the R-Studio:**

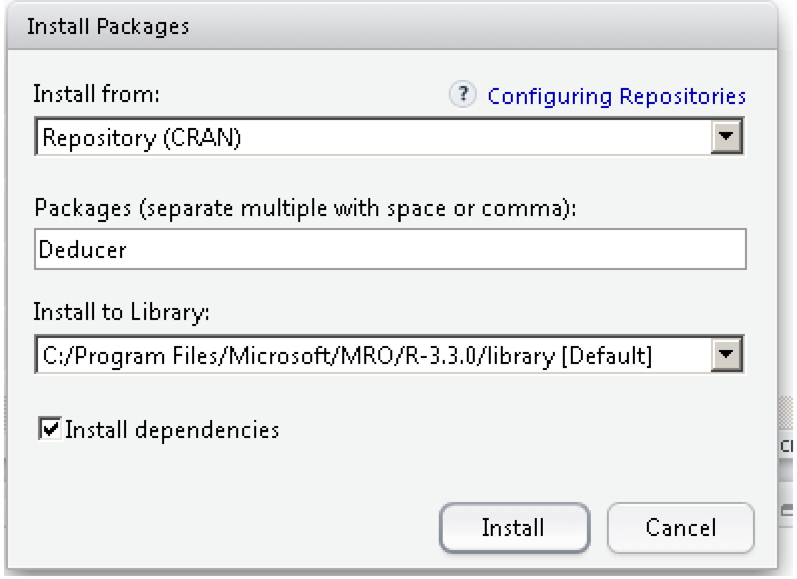
**(1) Set your work path for the Session (and current sessions)**

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"



Install from: Repository (CRAN)

Packages (separate multiple with space or comma): Deducer

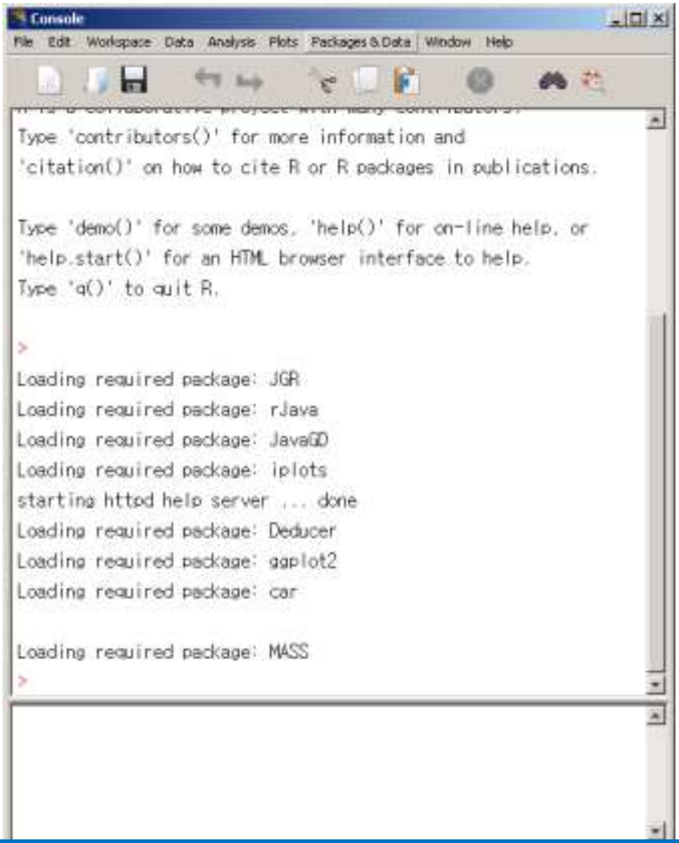
Install to Library: C:/Program Files/Microsoft/MRO/R-3.3.0/library [Default]

☑ Install dependencies

# Run Duducer

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

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

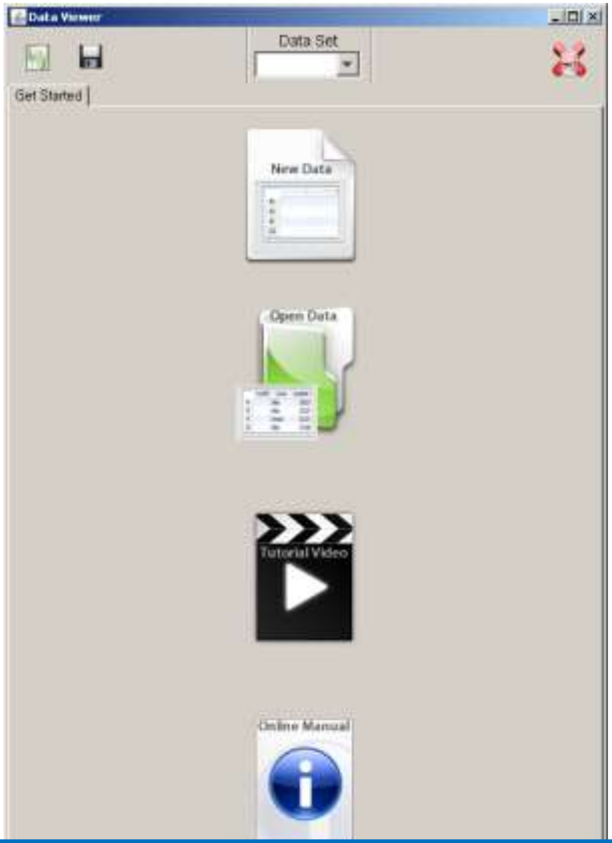


```
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

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.

>
Loading required package: JGR
Loading required package: rJava
Loading required package: JavaGD
Loading required package: iplots
starting httpd help server ... done
Loading required package: Deducer
Loading required package: ggplot2
Loading required package: car

Loading required package: MASS
>
```



# 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

# Introduction to Statistics

!