SOC6302 Statistics for Sociologists

Monica Alexander

Week 1: Introduction

Overview of today

- Overview of course
- Why learn statistics?
- Review concepts
- R, RStudio, Quarto

Overview of course

Instructor and TA

Instructor: Monica Alexander (she/her)

- Email: monica.alexander@utoronto.ca
- Office hours: Fridays 2-3pm, 700 University Avenue Level 9 Room 9135

TA: Aida Parnia (she/her)

- Email: a.parnia@utoronto.ca
- Office hours: TBA

Objectives

This course introduces statistical techniques and methods to analyze data to draw inferences about social processes. You will learn

- How to read in, describe, plot and analyze data in a statistical software that uses a programming language (R)
- Some important methods of statistical analysis to explore relationships between social phenomena
- How to communicate statistical results and limitations of your analysis

Objectives

Practical objectives:

- Getting more comfortable with data
- Getting more comfortable with coding
- Reading in data, data exploration, visualization, modeling
- Statistical literacy

This course will be very hands-on with coding and data manipulation.

Textbooks and other resources

There is no required textbook for this class. Some resources that might be useful:

- Diez, David, Cetinkaya-Rundal, Mine, and Barr, Christopher. 2019. 'OpenIntro Statistics' https://www.openintro.org/book/os/. pdf is free online.
- Field, Andy, Miles, Jeremy, and Field, Zoe. 2012. 'Discovering Statistics Using R'. One copy is available in the library; electronic copy can be purchased for ~$85.
- Alexander, Rohan. 2022. 'Telling Stories with Data' (TSWD) (https://www.tellingstorieswithdata.com/). Available freely online.

R Resources:

- TSWD
- Grolemund, Garrett and Wickham, Hadley. 2020. 'R for Data Science' (https://r4ds.had.co.nz/).

Software

We will be using the programming language R in this course, through RStudio.

- You most likely already have these installed
- More info on how to install these on Quercus
- Emphasis on `tidyverse` and Quarto

Assessment

- Three assignments (3x15% = 45%)
- Mid-term (20%)
- Research project (35%)

Assignments

- Data analysis with R
- Interpretation
- Hand in code, instructor/TA should be able to run without errors

Mid-term

- In class, after reading week
- Assess everything in Weeks 1-6
- Mostly short answer questions, some multiple choice
- No coding required but may need to interpret R code

Research Project

Choose a research question based on dataset(s) provided (or one of your choice), and perform statistical analysis

In the research project you will

- Develop a research question based on data set of choice
- Analyze data using methods learned in class
- Present, interpret and summarize findings

Research Project

Worth a total of 35%, but will be graded in two parts:

1. Research question and brief exploratory data analysis (10%)
2. Final report, which incorporates 1 (25%), due at end of semester.

Lecture + Lab format

- We have a three hour block, first 1.5-2 hours will be lecture style
- The rest will be a computer lab, working through exercises in R
- So bring your laptop!

Course Policies

- **Communication**: First, see if you can answer your question by checking the syllabus. Second, try to ask questions during class, tutorials, or office hours. Third, there will be a discussion board on Quercus. Fourth, email myself or your TA (please include the course number in the subject line)
- **Accessibility**: visit http://studentlife.utoronto.ca/accessibility as soon as possible.

Why learn statistics?

Why learn statistics?

As sociologists, we are trying to understand different aspects of society.

Statistical techniques give us a means to investigate and test research questions and policy impacts across different areas of people's lives.

Example research questions could include

- How is population mobility changing in the era of Covid-19?
- How do people cope with financial hardship?
- How does paid maternity leave affect women's workforce participation?
- Does volunteering increase your sense of wellbeing?

Why learn statistics

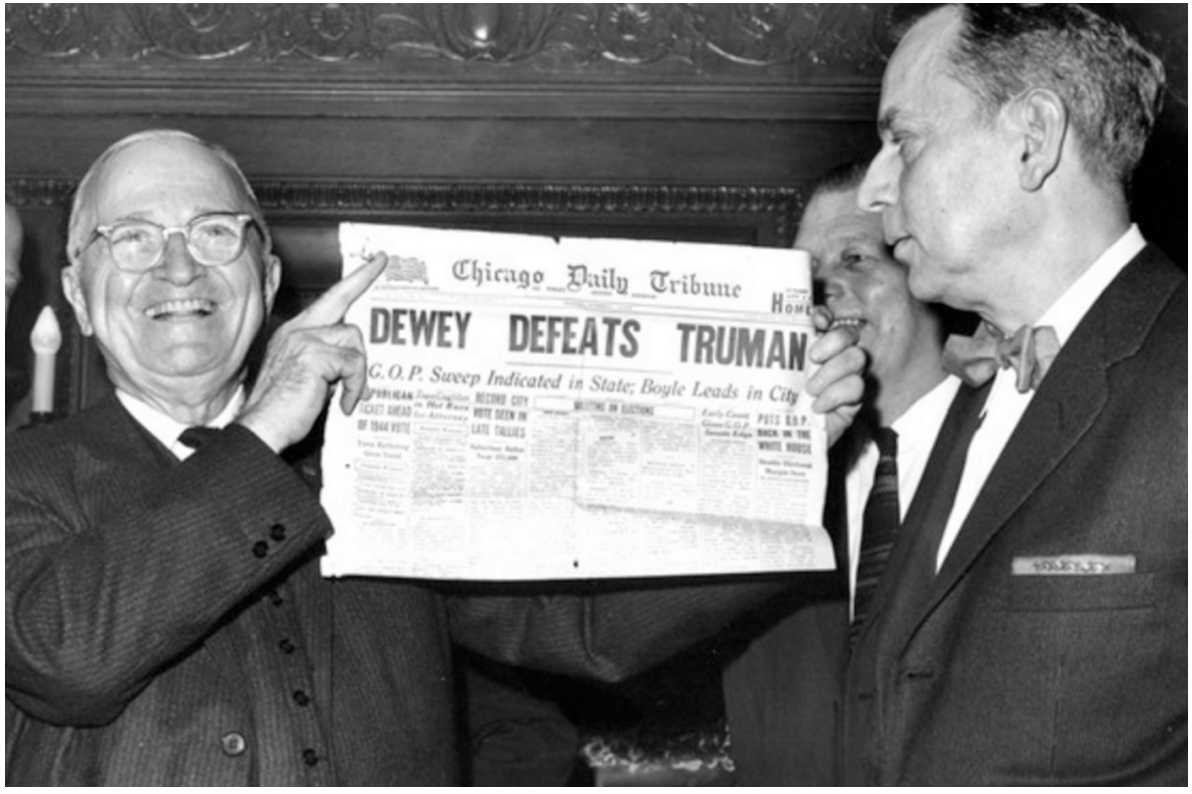
It's not just learning what you could do with data, it's learning what not to do with data

- How biases and selection can give misleading conclusions
- When is it inappropriate to use certain techniques

It's not just to support your own arguments, it's learning how to assess other people's arguments

- Statistics, data analysis and visualization is an art form
- Cutting through the lies, damned lies and (misused) statistics

Misleading statistics

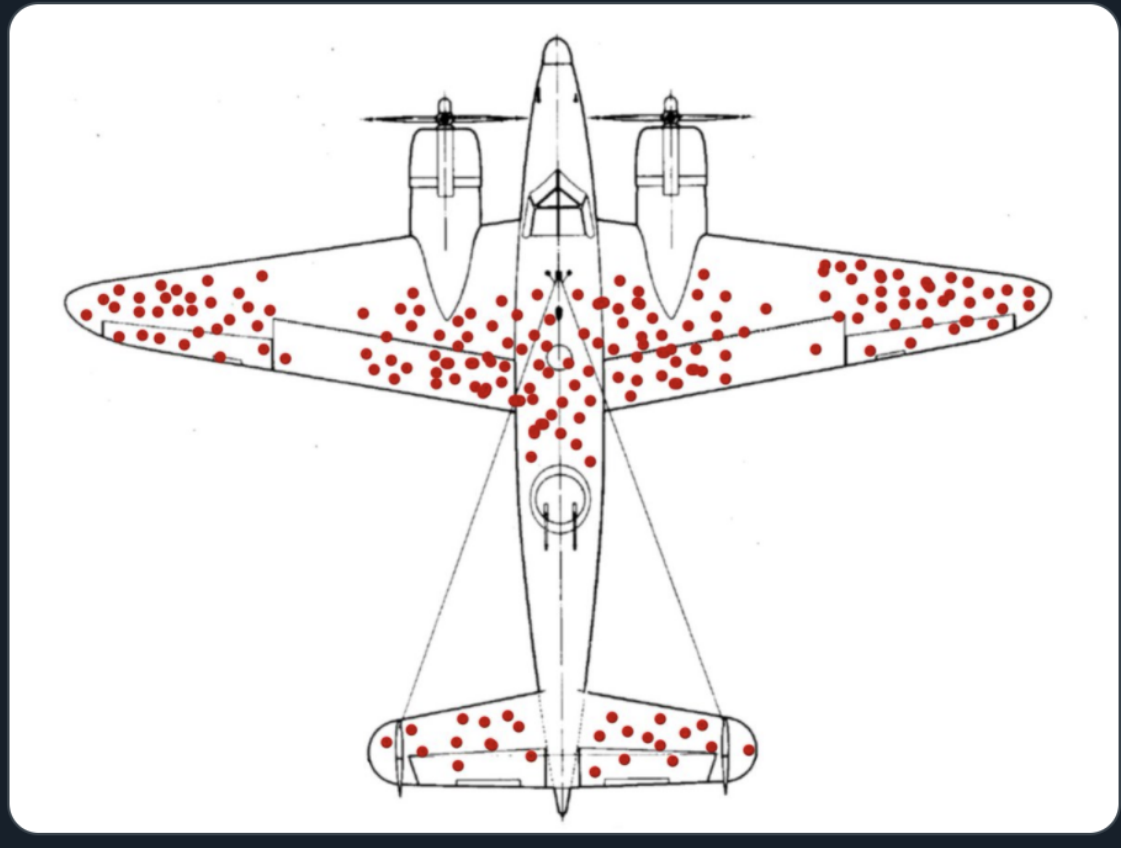


- Truman won the Presidential election in 1948
- This is a photo of Truman holding an up an erroneous headline
- Based on phone survey which predicted overwhelming win for Dewey
- What went wrong?

Bad stats becomes a meme

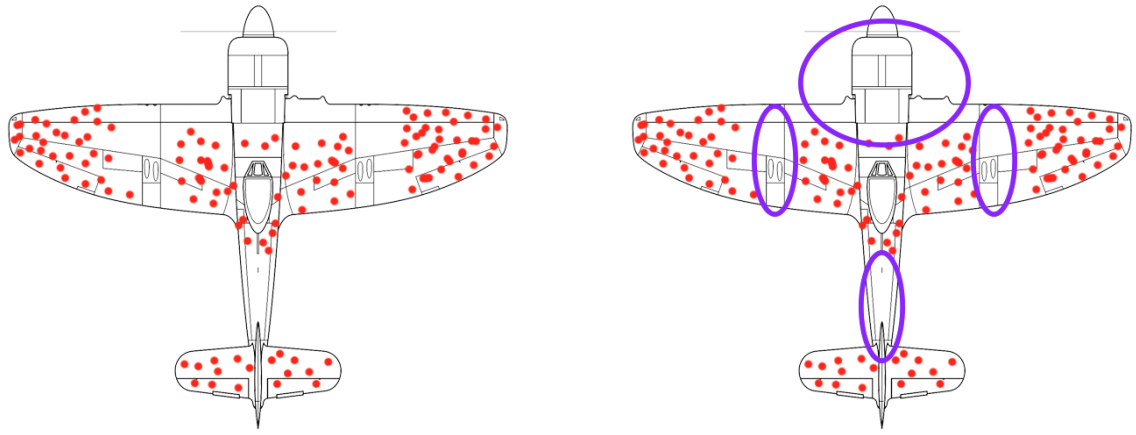
Health Nerd ✓
@GidMK

"I got COVID-19 and it was fine"



4:10 AM · May 1, 2021 · Twitter for Android

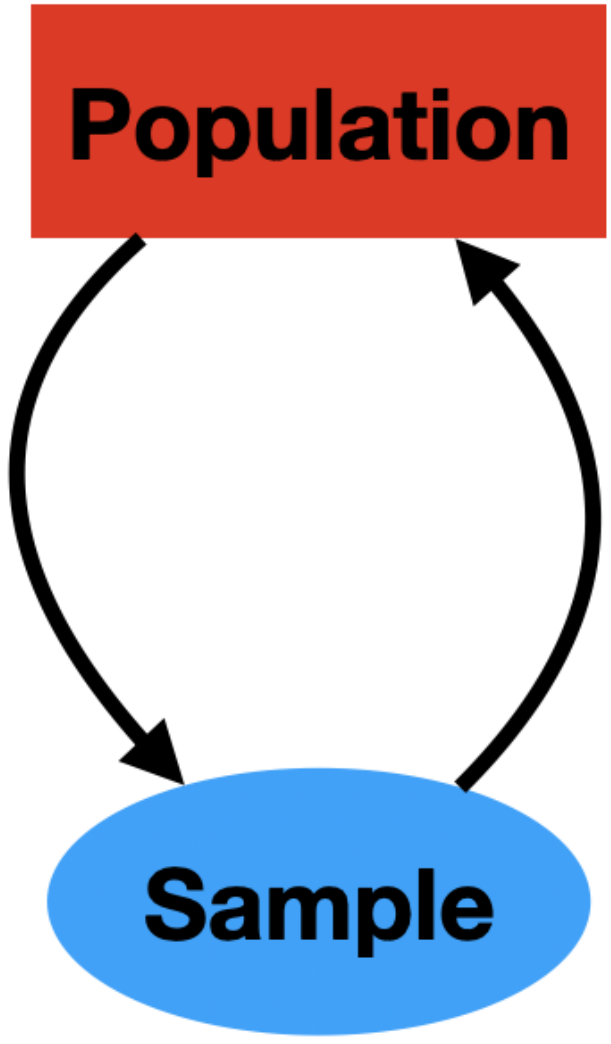
Misleading statistics



- Abraham Wald in WWII
- Want to place armor on planes in most effective place
- Gathered data from planes returning from battle and observed bullet holes
- Most holes in the fuselage, not so many in the engines
- Where should armor go?
- Can you think of other examples?

Review

What it all comes down to



Populations

At the core of statistical methods is wanting to say something about a **population** of interest.

What is a population? Depends on the context of study

- Everyone enrolled in university in Canada
- Everyone at UofT
- Everyone studying graduate-level sociology at UofT
- Everyone in this class

Samples

Say we want to study the relationship between hours studied and job placement for all graduate university students in Canada.

- Not really plausible to get data on this for the whole of Canada.
- In reality, we would collect data on a **subset** or **sample** of the population and try and generalize to the whole of Canada.
- With statistics, we are going to make conclusions based on what we see in the sample that we hope will be true for the population.

Samples

Our example: the relationship between hours studied and job placement for all university students in Canada.

- We could plausibly measure the hours studied and job placement for those student who took SOC252
- This class would be a sample of the population of interest, because you are all graduate university students in Canada.

Is it a good sample? What do I mean by good?

Sampling techniques

Terminology:

- **Element**: An element is an object or case, the unit on which a measurement is made.
- **Population**: The population is a collection of elements about which we wish to make an inference
- **Sampling units**: The sampling units are non-overlapping collections of elements in the population.
- **Sampling frame**: The sampling frame is a list of the elements or, for more complex samples, a list of the sampling units.
- **Sample**: A sample is a subset of the elements drawn from the population using one of several sampling methods.

What are each of these in our example?

Sampling techniques

Include:

- **Simple Random Sampling (SRS)**: A random sample is sometimes defined as a sample in which all possible elements have an equal chance of occurring.
- **Stratified Sampling**: based on variable of interest
- **Cluster Sampling**: SRS within clusters (e.g. districts within a province, schools within districts)
- **Convenience Sampling**: as the name suggests, easy to get, with no consideration of how sample was drawn from broader population

What is the sampling method in our example?

Two main domains of statistics

- **Descriptive statistics**: uses the data to provide descriptions of the population, either through numerical calculations or graphs or tables.
- **Inferential statistics**: makes inferences and predictions about a population based on a sample of data taken from the population in question.

We will cover both types in this course. Understanding patterns in descriptive statistics is essential to doing good inferential statistics.

Variables

Traits, characteristics, outcomes that we are interested in. e.g.

- hours of study
- course grade
- industry of job placement
- province of residence
- age
- self-reported health

Variables

Often we are interested in studying the relationship between two or more variables.

- The **outcome** of interest is the **dependent variable**
- Variables **used to explain the outcome** can be called
  - independent variables
  - explanatory variables
  - covariates
  - predictors

I will use these terms interchangeably.

What are the independent and dependent variables in our example?

Types of measurement of variables

- **Quantitative**: has a numeric meaning
  - **Continuous**: any possible number
  - **Discrete**: possible values can assume only certain values, usually the counting numbers
- **Qualitative**: categorical, no numeric meaning

What are the types of variables in our example?

Random variables

- A **random variable** is a variable whose values depend on the outcomes of a random process.
- For our purposes, the "random process" is taking a random sample of a population
- For example, consider the variable annual income:
  - We randomly select someone from the population and note their income.
  - The value of this depends on the person who was selected
  - If we randomly selected someone else again, it's likely that the income value would be different

RVs are the basis of probability and statistical inference! The randomness is what we need to quantify, to separate the signal from the noise.

Summary measures of quantitative data

Summary measures of quantitative data

Pretend you have a set of observations of a quantitative variable, e.g. everyone's height in this class. We often want to **summarize** our set of observations with one or more numbers. Often interested in:

- Measures of **central tendency**, i.e. what would we expect someone's height to be, what's the most common height
- Measures of **spread**, i.e. what are the ranges of heights observed, what is the deviation of heights away from the expected height?

Some common symbols and notation

There are some common notation that will come up over and over. To start with

- Population size $= N$
- Sample size $= n$
- A particular individual in a sample denoted by index i
- Random variables (note these are captials!):
  - Dependent variable: Y
  - Independent variables X
- A set of random variables for individuals $i = 1, 2, \ldots, n$: $X_1, X_2, \ldots, X_n$
- Specific values or outcomes of the corresponding random variables:
  - Dependent variable: y
  - Independent variables x

Measures of central tendency

- **Mean**: the average
  - Population mean usually denoted as μ
  - Sample mean denoted with a bar e.g. $\bar{x}$

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

- **Median**: the value for which 50% of the sample is below and 50% of the sample is above. It is the 50% percentile. To calculate
  - Order set of values from smallest to largest
  - find the middle number
- **Mode**: the value that occurs the most frequently

Measures of variability

- **Range**: The difference between the minimum and maximum value
- **Interquartile range**: The difference between the 25% and 75% percentiles. To calculate
  - Order set of values from smallest to largest
  - Separate into quarters
  - Find the first quarter (Q1) and third quarter (Q3)
  - IQR = Q3 - Q1

Measures of variability

- **Variance**: average of the squares of the deviations
  - Population variance: σ^2
  - Sample variance: s^2

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

- **Standard deviation**: average of the deviations
  - Population standard deviation: σ
  - Sample standard deviation: s

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} = \sqrt{s^2}$$

Introduction to R and Quarto

R and RStudio

- You will need to download and install both R and RStudio
- More info on Quercus
- Please do this as soon as possible if you haven't already

This is review from bootcamp! See those materials for more revision: https://www.monicaalexander.com/teaching/

What is R?

- R is a programming language for statistical computing and graphics
- Using R is like speaking another language (but you type it)

You may have used other programs to do statistical calculations before (Excel, SPSS, Stata, SAS)

- With R you have to give the computer typed commands in order for it to do stats (rather than clicking buttons)
- Much more powerful methods

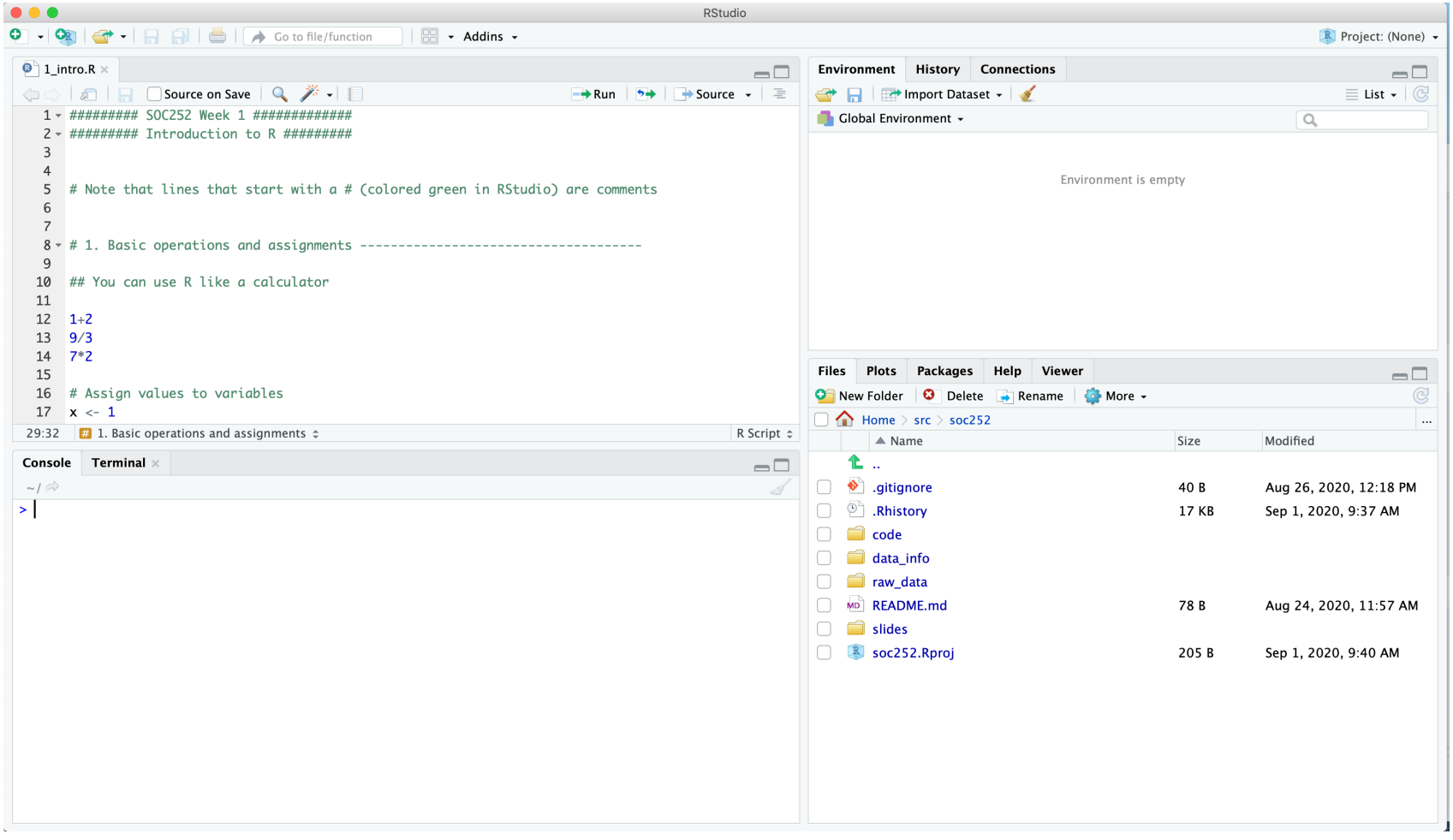
What is RStudio?

- RStudio is an integrated development environment for R.
- It makes it easier to write R code and visualize inputs and outputs.
- You will need to download and install both R and RStudio

Think of a car analogy:

- R is the engine
- RStudio is the car dashboard (steering wheel, controls etc)

RStudio


RStudio
Go to file/function
Addins
Project: (None)
1_intro.R
Source on Save
Run
Source
1 ######### SOC252 Week 1 #############
2 ######### Introduction to R #########
3
4
5 # Note that lines that start with a # (colored green in RStudio) are comments
6
7
8 # 1. Basic operations and assignments ----------------------------------
9
10 ## You can use R like a calculator
11
12 1+2
13 9/3
14 7*2
15
16 # Assign values to variables
17 x <- 1
29:32
1. Basic operations and assignments
R Script
Console
Terminal
~/
>
Environment
History
Connections
Import Dataset
List
Global Environment
Environment is empty
Files
Plots
Packages
Help
Viewer
New Folder
Delete
Rename
More
Home
src
soc252
Name
Size
Modified
..
.gitignore
40 B
Aug 26, 2020, 12:18 PM
.Rhistory
17 KB
Sep 1, 2020, 9:37 AM
code
data_info
raw_data
README.md
78 B
Aug 24, 2020, 11:57 AM
slides
soc252.Rproj
205 B
Sep 1, 2020, 9:40 AM

Writing code in R

1. R Console: Executes each line of code as you go; does not save code for later use
2. R Script: Saves code and comments in a file so you can select some or all of the code in a script file to run; does not include output
3. R Markdown or **Quarto**: A file which combines text and chunks of R code (which can be executed independently). This allows you to see output without "knitting" the whole file. Quarto is the newer version of R Markdown.

We will focus on number 3.

Quarto documents

Source Visual

```
---
title: "Example Quarto"
author: "Monica Alexander"
format: pdf
editor: visual
---

## Introduction

Here is some text

## Analysis

Here is some code

```{r}
2+2
```
```

Knitted to a PDF

Example Quarto

Monica Alexander

Introduction

Here is some text

Analysis

Here is some code

```
2+2
```

```
[1] 4
```

R code review

R as a calculator versus defining objects

```
2+87
```

```
## [1] 89
```

`x` is assigned the value of 7:

```
x <- 7
x
```

```
## [1] 7
```

Types of variables

- numeric
- logical
- character
- factor

```
x <- 7
is_this_soc252 <- TRUE
my_name <- "Monica"
fruit <- as.factor("banana")
```

R Packages

- A lot of people have written **R packages**, which are add ons to base R that increase the functionality

Think of a phone analogy:

- R/RStudio is a phone
- R packages are apps

We will be using a few different R packages quite a lot during the course, e.g.

- `dplyr` (data manipulation)
- `ggplot2` (graphing)

These and other packages can be downloaded through downloading the `tidyverse` package (you probably already have this installed)

Different types of objects in R

- single values
- vectors
  - contain two or more values
  - Defined with the c() function ("concatenate")
  - Values must be of the same type
- data frames or tibbles
  - Closest thing to a dataset that we deal with
  - Each column is a different variable, each row is an observation
  - Columns (variables) can be different types

Different types of objects in R

```r
# single value
y <- 2
# vector
my_numbers <- c(7,4,3,2)
my_names <- c("Monica", "Rohan")

y
```

```
## [1] 2
```

```r
my_names
```

```
## [1] "Monica" "Rohan"
```

Functions

- Do stuff to your variables!
- Have already seen some: as.factor(), c()
- Examples:
  - mean(), median()
  - min(), max()
  - length(), dim()
  - paste()
  - is.numeric() etc

Functions

Create a vector of numbers:

```
x <- c(1,3,5,7,9)
```

Calculate summary statistics:

```
mean(x)
```

```
## [1] 5
```

```
max(x)
```

```
## [1] 9
```

```
IQR(x)
```

```
## [1] 4
```

Where to get help

- Intro to R:
  - R4DS is the most relevant textbook for learning "tidyverse" R
  - Telling stories with data
- Lab time and office hours
- Google, google, google
  - Don't expect you will be able to code for memory to start off with
  - Think about what you want to do and then if you don't know how to do it, Google key terms
  - Googling errors is also very helpful