Essential Statistical Concepts and Vocabulary

School of Information Studies
Syracuse University

Learning Topics for This Week

Explain the purpose of inferential statistics.

Inventory the math skills needed for this course.

Install or upgrade R and RStudio if you have not done so.

Run your first command in R.

Memorize definitions of key concepts in descriptive statistics.

Explain the connection between deviations from the mean, sum of squares, variance, and standard deviation.

Use R to calculate the standard deviation.

Identify shapes of certain basic distributions.

Use R to generate histograms that illustrate distributions.

The Purpose of Inferential Statistics

William Sealy Gosset

1876-1937

Chemist, researcher, statistician

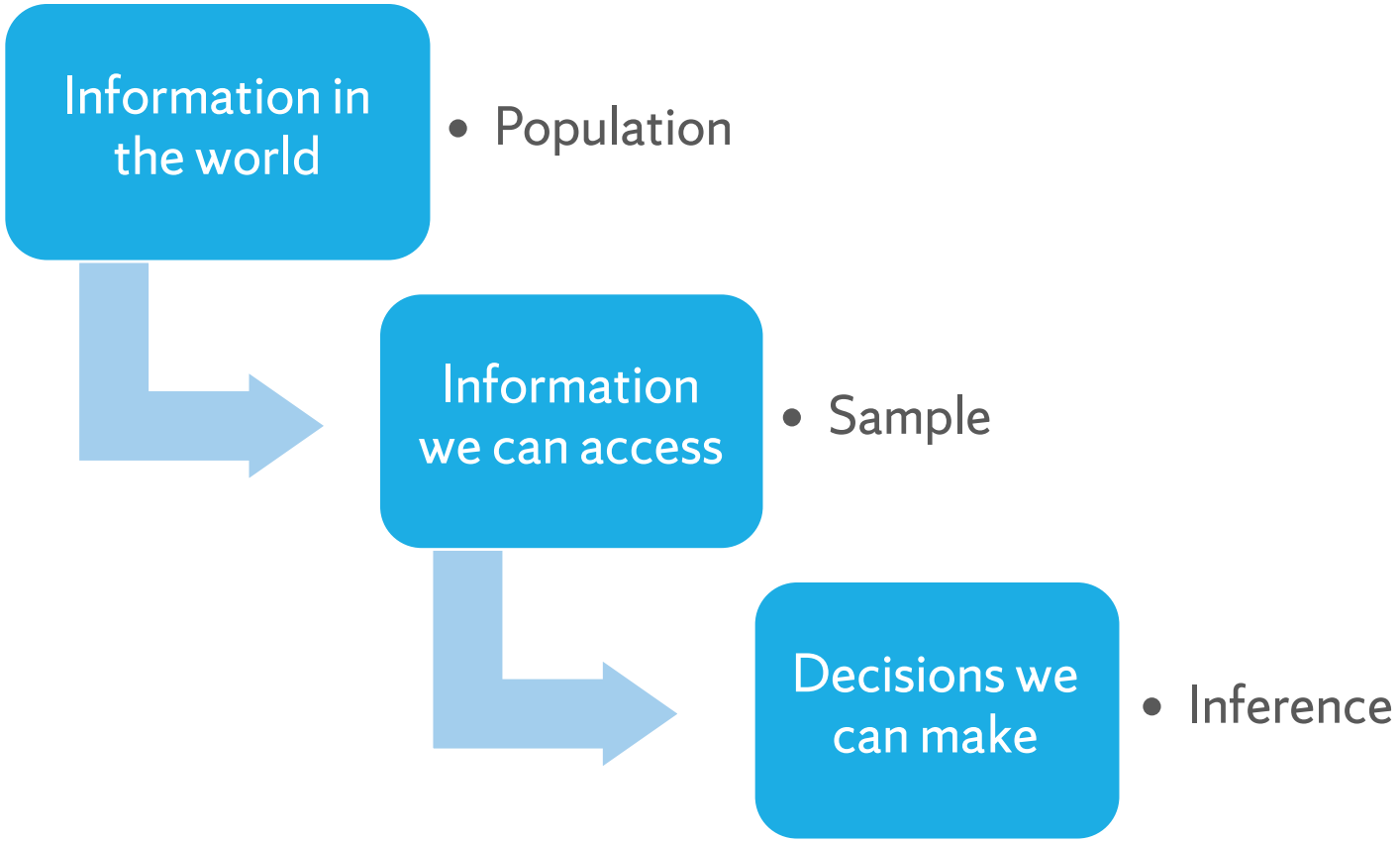
Worked at Guinness Brewery

Used the pseudonym “Student”

Developed “Student’s t-test”

Image credit: User Wujaszek on pl.wikipedia - scanned from Gosset’s obituary, Public Domain, https://commons.wikimedia.org/w/index.php?curid=1173662

Making Inferences



Decisions Under Uncertainty

99.9% correct decisions

0.1% incorrect decisions

Math Skills

Math Skills and More

Add, subtract, multiply, and divide, on paper or with a calculator.

Follow the meaning and usage of algebraic equations such as $y = 2x - 10$.

Organize and review columns and rows of data, as one would typically find in a spreadsheet.

Interpret basic graphs, such as bar charts and scatterplots.

Basic Math in R

```
> 2 + 2
[1] 4
> 4 * 4
[1] 16
> 8/8
[1] 1
> (4 * 4) - (8/8)
[1] 15
```

Basic Equations in R

```
> x <- 10
> y <- 2 * x - 10
> y
[1] 10
```

Rows and Columns in R

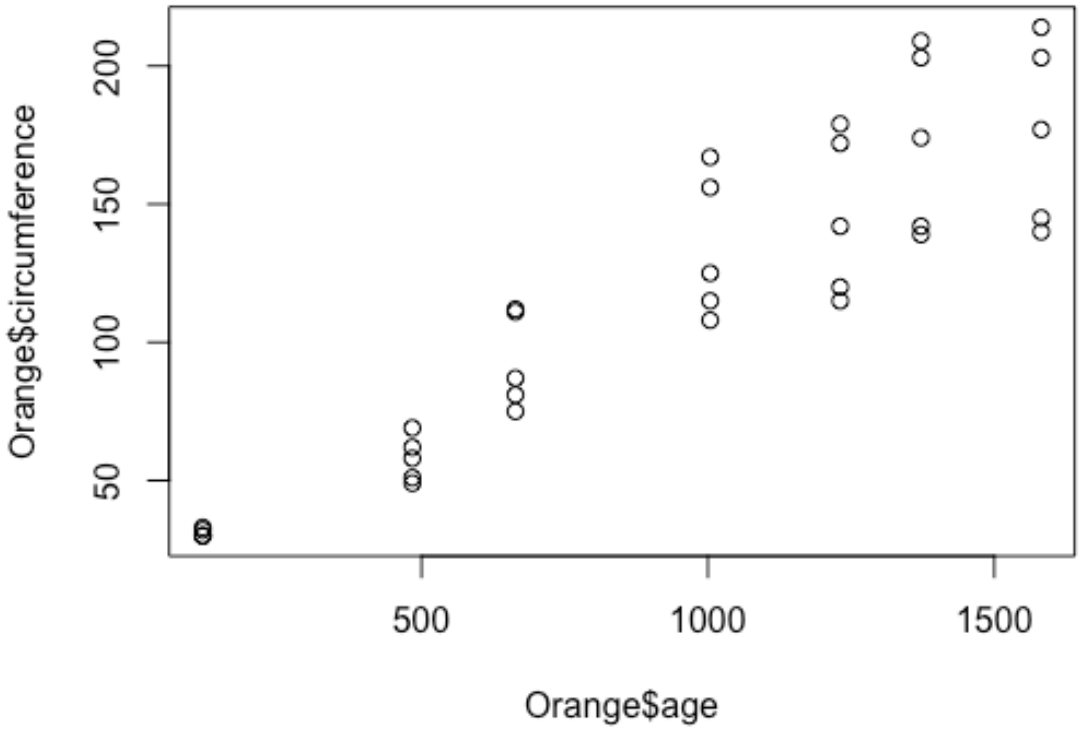
> View (Orange)

	Tree	age	circumference
1	1	118	30
2	1	484	58
3	1	664	87
4	1	1004	115
5	1	1231	120
6	1	1372	142
7	1	1582	145
8	2	118	33
9	2	484	69
10	2	664	111

Showing 1 to 11 of 35 entries

Basic Graphs in R

> Plot (Orange$age,Orange$circumference)



Install/Upgrade R and RStudio

CRAN: Comprehensive R Archive Network

At https://cran.r-project.org, you will find links to download and install R. Make sure to install R first, before installing Rstudio.

The Comprehensive R Archive Network

Download and Install R

Precompiled binary distributions of the base system and contributed packages, **Windows and Mac** users most likely want one of these versions of R:

- 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 your Linux package management system in addition to the link above.

Mirrors

The CRAN webpage may ask you to choose a “mirror.”

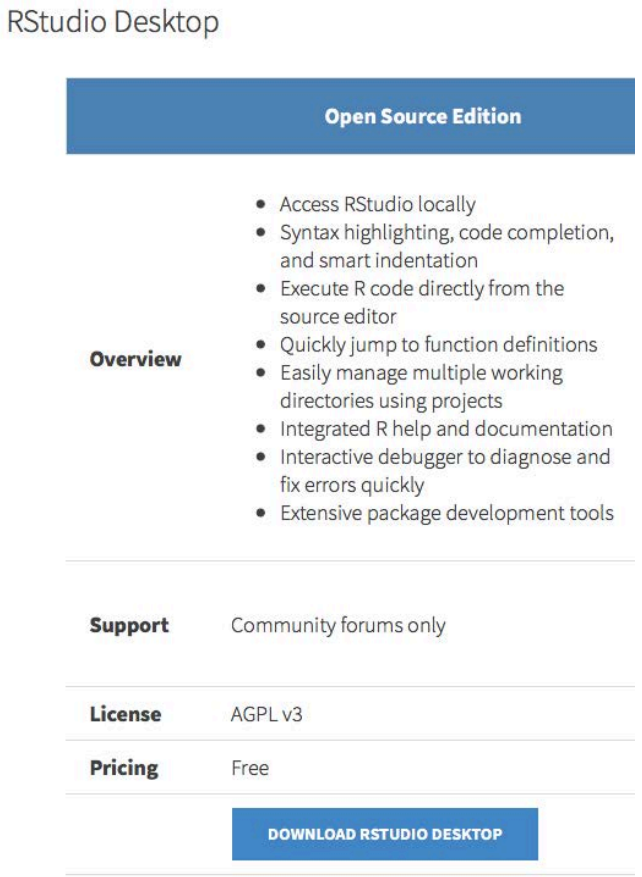
A mirror is simply a copy of a website that is hosted somewhere else.

As an open source, noncommercial package R is mirrored at many different locations.

Choose the mirror that is geographically closest to you.

RStudio

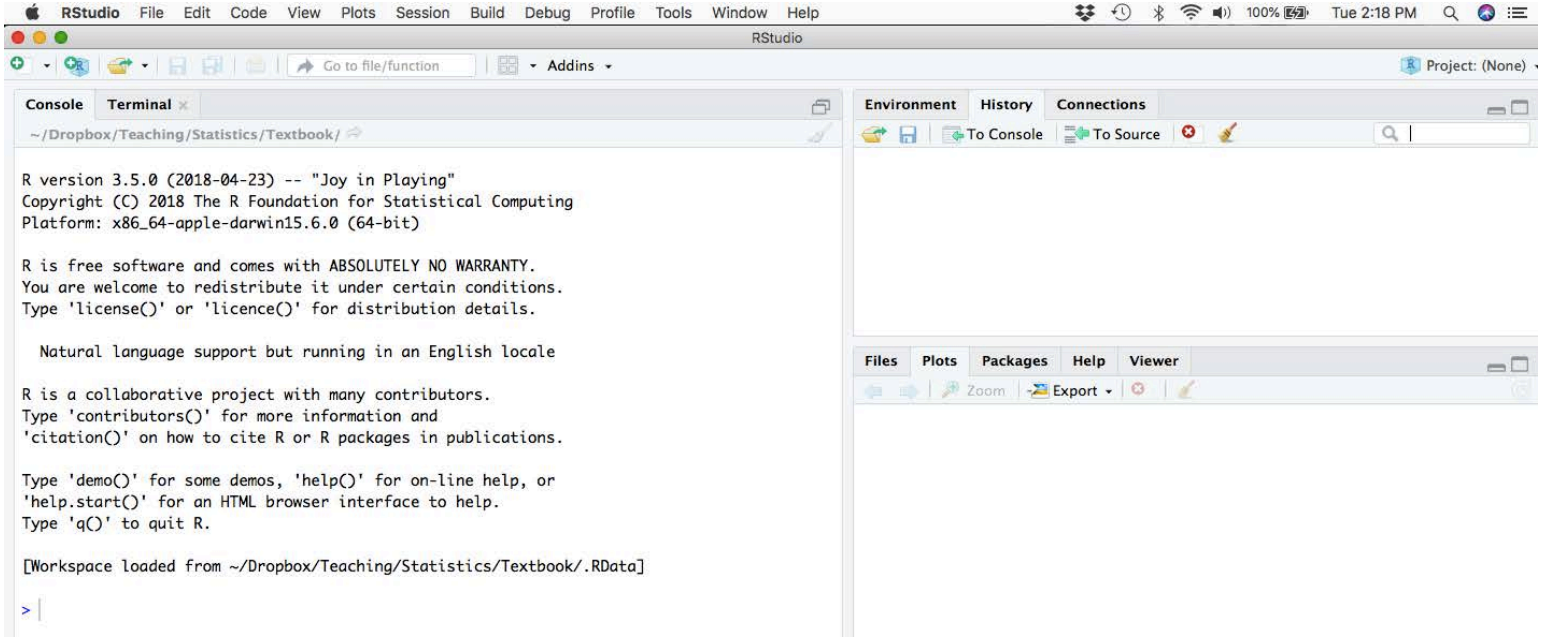
At https://www.rstudio.com/products/RStudio/#Desktop, you will find links to download and install RStudio. You should download the free desktop version of RStudio.



Run Your First Command in R

How to Run Your First Command

The RStudio display is divided into panes. The "console" is on the lower left and it is where you run commands.



The c() Command: Combine Elements Into a Vector or List

Let's say we have seven values representing the number of times a user clicks "Next" on a multipage website:

0, 0, 0, 2, 3, 4, 5

To combine these values into a "vector" (a list of elements, all of the same type) use the c() command:

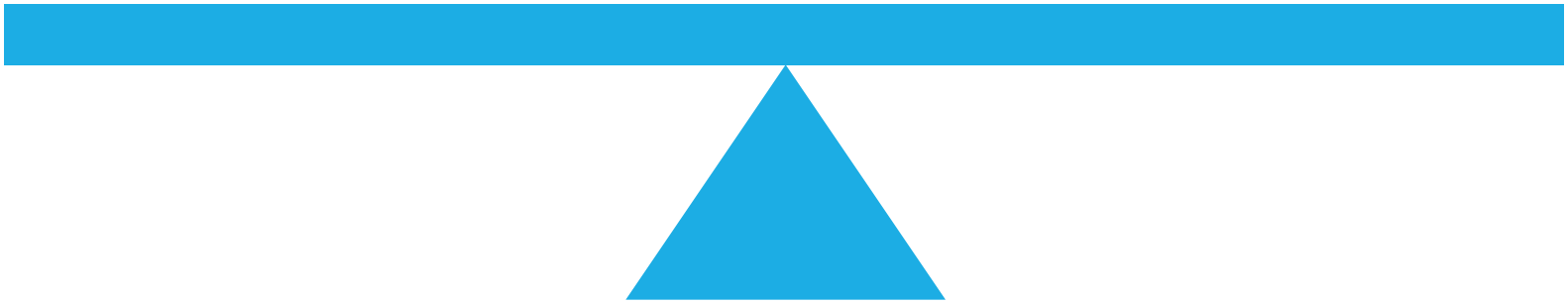
c(0,0,0,2,3,4,5)

R will respond by "echoing" the vector to the console

Key Concepts in Descriptive Statistics

Mean, Median, and Mode

0, 0, 0, 2, 3, 4, 5



Why Measures of Dispersion?

Central tendency is an important characteristic of a collection of numeric data.

Answers the question: What is typical?

Another aspect that matters a lot is the “spread” or “dispersion” of the data.

Imagine if we were filling trucks with freight and we needed to choose between two types of trucks: 7 tons or 10 tons.

Containers in group A are 2, 5, and 5 tons (mean is 4 tons).

Containers in group B are 1, 3, and 8 tons (mean is 4 tons).

The Connections Among Measures of Variability

Connections Between Measures of Dispersion

1. **Range:** the difference between the minimum and maximum observation
2. **Deviations from the mean:** a signed list of differences between each observation and the mean
3. **Sum-of-squares:** the sum of the squared deviations from the mean
4. **Variance:** the average "squared deviation" from the mean
5. **Standard deviation:** the square root of the variance

Items 2 and 3 are steps along the way to calculate 4 and 5

Which Works Better?

The variance among heights of those school children was 14 squared inches.

The standard deviation among heights of those school children was 3.7 inches.

The variance of weights of those shipments was 9 squared tons.

The standard deviation of weights among those shipments was 3 tons.

Use R to Calculate Standard Deviation

Identify Shapes of Distributions