

Overview

Introductions

Overview and logistics of the course

Review of a few central concepts from Intro Stats

Introduction to R

- R as a calculator
- Objects and vectors
- Installing the class SDS230 package and LaTeX (if there is time)



Office hours and contact information

Ethan Meyers (he/him)

Email: ethan.meyers@yale.edu

Office hours:

- Tuesdays and Thursdays, 10:45-11:45
 - (subject to change)

Office: Kline Tower, room 1253

https://yale.zoom.us/j/95781950309

Note: I'm going to be a little distracted this semester by a new neural network...

Teaching Assistants

Teaching Fellows (TF)

- Ashley Oaks <u>ashley.oaks@yale.edu</u>
- Linghai Liu <u>linghai.liu@yale.edu</u>
- Yunzhe Jiang <u>yunzhe.jiang@yale.edu</u>
- Ruiqi Li <u>ruiqi.li@yale.edu</u>

Undergraduate Learning Assistants (ULA)

- Sohum Kapadia <u>kapadia@yale.edu</u>
- Tejas Muthusamy <u>tejas.muthusamy@yale.edu</u>
- Christian Baca <u>Christian.Baca@yale.edu</u>

Course manager

Huaze Gao <u>huaze.gao@yale.edu</u>



Introductions

Let's do some quick introductions

Create groups of 3-4 people:

- Your name and preferred gender pronouns
- Your major/grad dept (research area)
- Why you are interested in this class
- Anything else you would like to share with your group

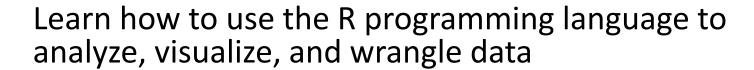




Course objectives

Solidify and extend concepts and method learned in intro stats

- Permutation tests, multiple regression, etc.
- Focus on intuition of why methods work rather than proofs



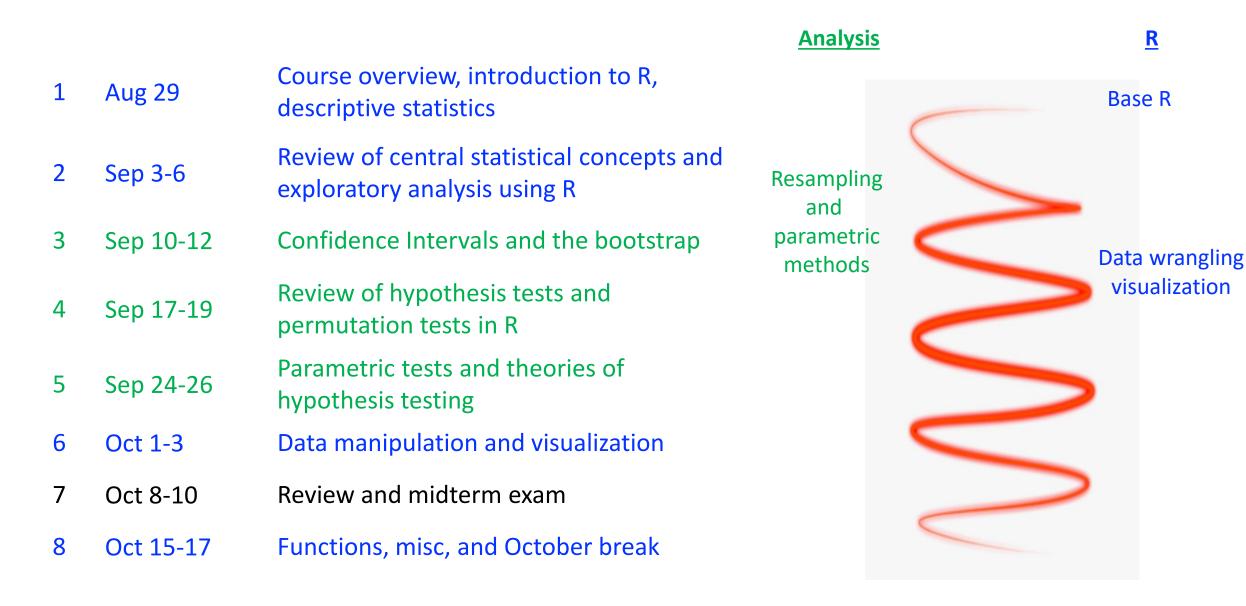
Gain experience extracting insights from real data

Learn how to find patterns in a large noisy data sets and convincingly convey the results to others!

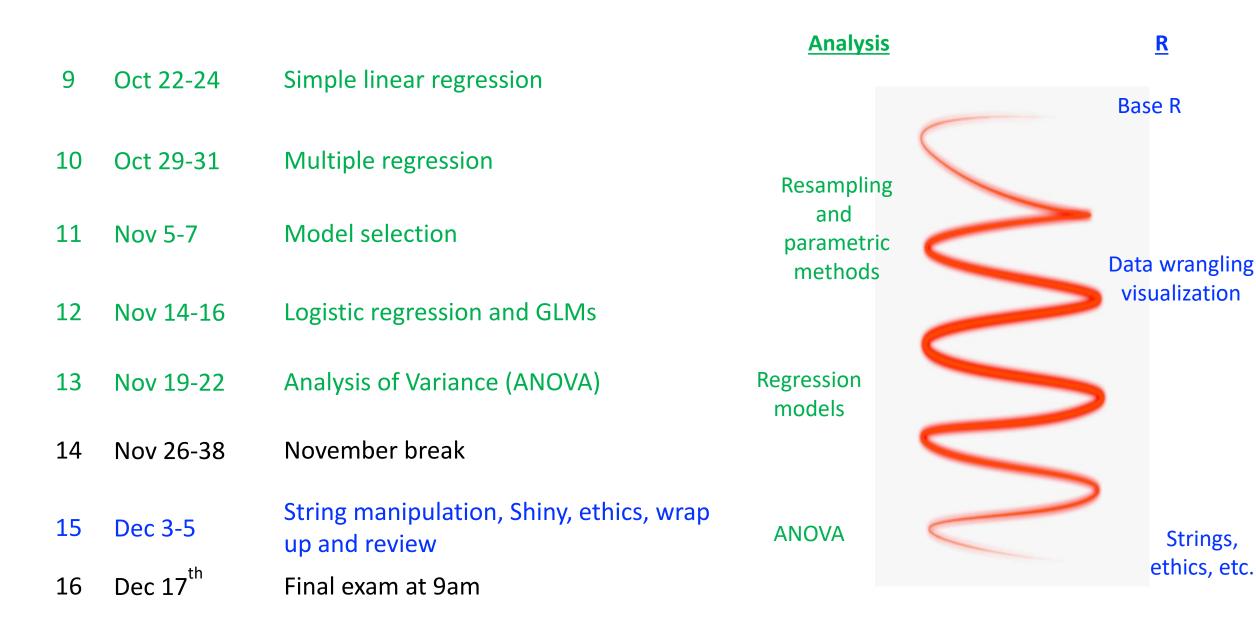




Plan for the semester



Plan for the semester



List of topics

R and descriptive statistics/plots: Base R, fundamental concepts in Statistics

Review confidence intervals: Sampling and bootstrap distributions

Review of hypothesis tests: Permutation and parametric tests, theories of testing

Data wrangling: filtering and summarizing data, joining data sets, reshaping data

Data visualization: grammar of graphics (mapping?)

Regression: simple/multiple, non-linear terms

ANOVA: one-way/factorial, interactions

Statistical learning: cross-validation, logistic regression (PCA, clustering?)

Examples of questions we might look at...

Bootstrap confidence intervals: How much do avocados typically cost?



ANOVA: Are all genres of movies rated the same on average?











Prerequisites

An introductory class in Statistics (AP or 10X)

 We will review Intro Stats concepts using computational methods, but we will be going through the material at a fast pace

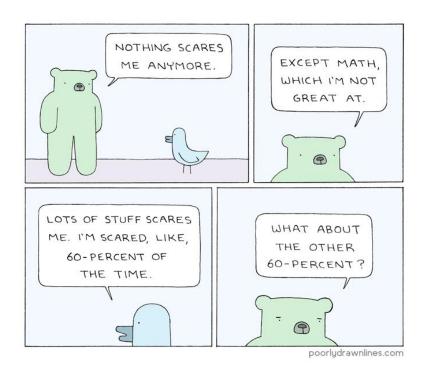
A large component of this class will be using the R programming

No prior programming experience needed!

Minimal mathematical prerequisites

 Many other S&DS classes to learn mathematical derivations of the methods we will use





Class logistics

Class time 9-10:15am Tuesdays and Thursdays

New content introduced, questions answered

Canvas website:

https://yale.instructure.com/courses/98725

No required textbook, reading resources will be posted to Canvas and in the homework assignments



Hadley Wickham & Garrett Grolemund

Office hours

My planned office hours (subject to change)

- Tuesday and Thursday at 10:45-11:45am
- Office hours will be on zoom and in Kline Tower room 1253

TA office hours are posted on calendar on Canvas

• We will try to have consistent office hours, although they might change particularly at the start of the semester

For questions about content covered in class, best to first ask on Ed Discussion

 Class participation grade based on questions and answers on Ed Discussion



Assignments and grades

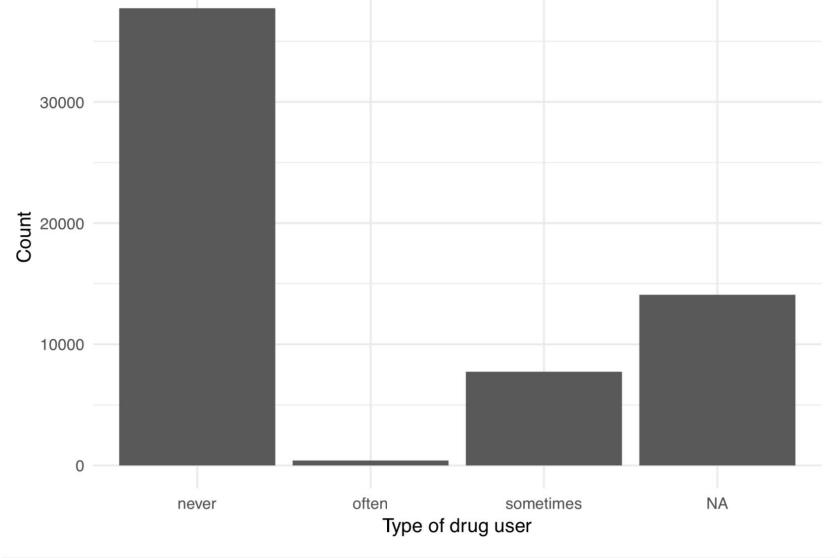
- 1. Homework problem sets (45%)
 - Exploring concepts and analyzing data using R
 - Weekly: 10 total



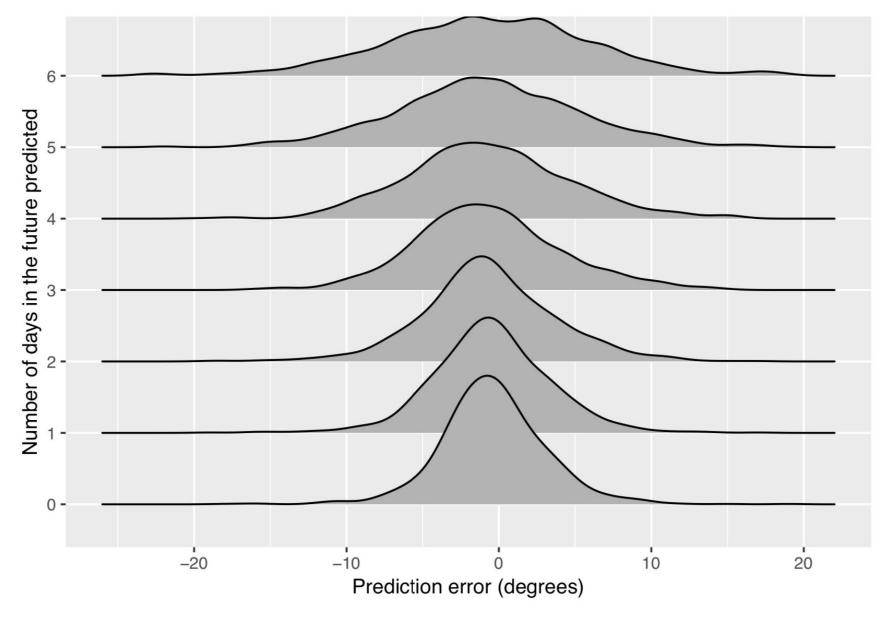
Homework policies

- You may discuss questions with other but the work you turn in must be your own
- Homework assigned on Tuesdays and are due at 11pm on Sundays
 - (with a 59 minute grace period)
- Late worksheets (90%) credit if turned in by 11:59pm on Monday
 - For any other extension a Dean's Extension is needed
- Lowest scoring homework will be dropped!

Example homework assignment piece



Example homework assignment piece



Answers: Personally I like the joy plot best here because it most clearly shows how the distribution becomes more spread out for predictions made further in the future (although all three plots do a reasonable job of showing this).

Assignments and grades

- 2. Final project (10%)
 - Find a data set and analyze it on your own (5-7 page report)

- 3. Exams (43% total)
 - Midterm (15%) Oct 10th during class
 - Final (28%) Dec 16th at 7pm
- 4. Participation (2%)
 - Active asking and answering questions on Ed Discussions
 - Full credit will be given for 8 or more questions or answers

Grade distribution

Grade cut-off are

- A [94-100], A- [90-94), B+ [87-90), B [80-84), etc.
 - I might slightly modify these downward if the class too hard

No strict grade distribution but roughly:

• 25% A, 25% A-, 25% B+, 25% everything else

Students generally score high on the homework (> 90) and exam scores tend to be lower ($^{\sim}80$)

If an exam is too hard, I sometimes curve them by adding "free points"

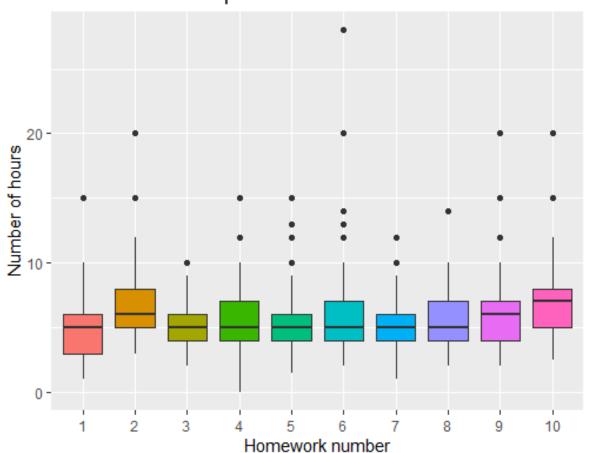
• E.g., if an exam is out of 85 points, I might add a free 15 bonus points so the exam is out of 100



Please try to focus on the learning rather than the grade!

How much work is this class?

Number of hours spent on homework



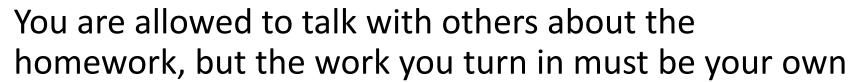
Relative to other courses you have taken at Yale, the workload of this course was:

Questio	n Average	Dept	Div		School
3.6 ±0.8	3	3.3 ±0.8	3.3 ±	±1.0	3.1 ±1.0
30 —					
25 —				n=28	
			n=24		
20 —					
15 —					
10 —					
					n=8
5 ——		n=5			
0 —	n=0 				
	much less	less	same	greater	much greater

Academic honesty

Plagiarism/cheating

Yale's Academic Integrity Statement



- Do not share answers
- Do not copy answers off the Internet
- Do not look at past year's homework





ChatGPT

Can use as a reference

- E.g., "What is the function to do x?"
 - i.e., ok to use it like Google/Stack Overflow

Do not use it to answer full questions

• i.e., do not type a homework question in chatGPT

To be an efficient programmer, it's important to be fluent with the material

 And if you don't learn the material, you will be in a lot of trouble on the exams



Class background survey

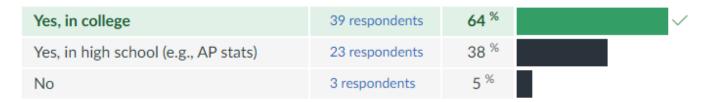
In order for me to get to know you and to better adjust the class to your interests, please fill out the <u>class background survey</u> on canvas

Under the Quizzes link on the left on Canvas

Class survey results from previous years

Have you taken an Introductory Statistics class before? Note: it is strongly recommended that you have before taking this class.

2023



Have you taken an Introductory Statistics class before?

2022

Yes, in college	37 respondents	59 %	~
Yes, in high school (e.g., AP stats)	27 respondents	43 %	
No	3 respondents	5 %	

Have you taken an introductory Statistics class before?

2021

Yes, in college	40 respondents	60 %		~
Yes, in high school (AP stats)	30 respondents	45 %	Correct Answer	
No	6 respondents	9 %		

Class survey results from previous years

Which Statistics methods/concepts are you comfortable with?

t-tests	41 respondents	67 %	✓
Confidence intervals	48 respondents	79 %	
The bootstrap	11 respondents	18 %	
Permutation tests	8 respondents	13 %	
One-way ANOVA	19 respondents	31 %	
Multiple regression	19 respondents	31 %	
Logistic regression	20 respondents	33 %	
Sampling distributions	32 respondents	52 %	
None of the above	8 respondents	13 %	

Class survey results from previous years

How much experience do you have with computer programming?

2023

Never programmed before	17 respondents	28 %	~
Some basic experience	31 respondents	51 %	
Intermediate	10 respondents	16 %	
Advanced	3 respondents	5 %	

2022

Never programmed before	9 respondents	14 %
Some basic experience	40 respondents	63 %
Intermediate	11 respondents	17 %
Advanced	3 respondents	5 %

2021

Never programmed before	17 respondents	25 %
Some basic experience	35 respondents	52 %
Intermediate	12 respondents	18 %
Advanced	3 respondents	4 %



Quick Review of central concepts in Intro Statistics

Yale Poorvu Center for Teaching and Learning

Top Ten Teaching Strategies

1. Learn every student's name.

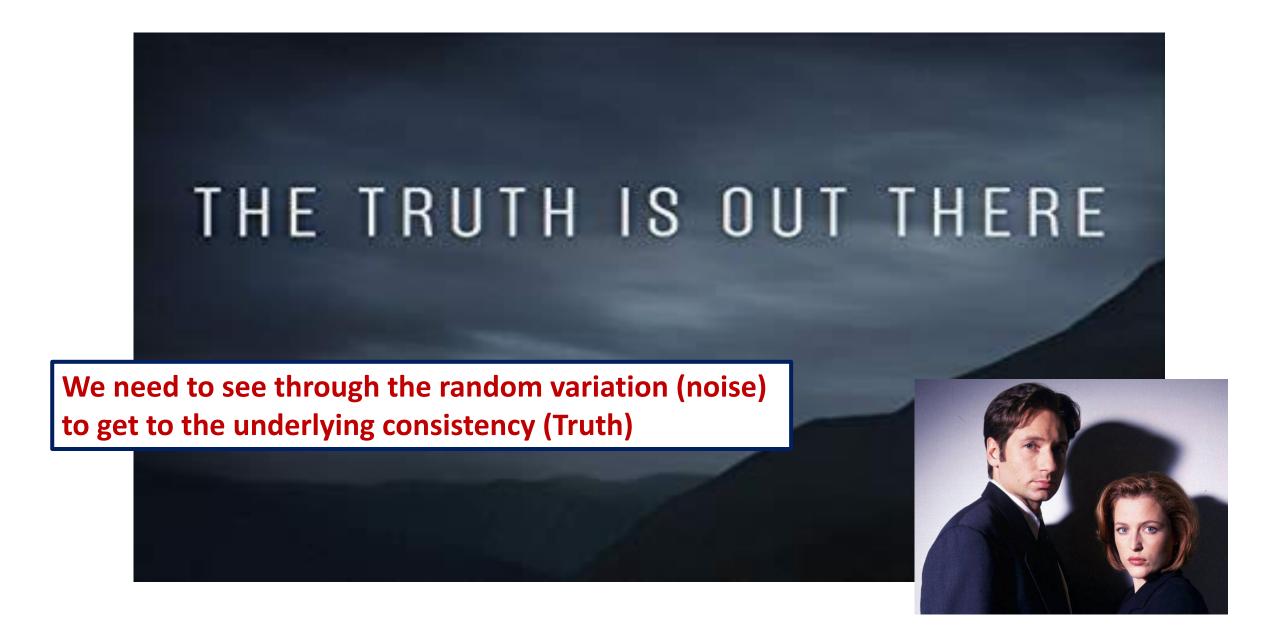
- 2. Create course objectives and classroom policies as a way to begin establishing community, and review them at midterm or more, as needed. In addition, discuss each session's learning objectives in class, with each meeting. Being explicit about your pedagogical techniques helps students see the design behind their learning.
- 3. Identify and utilize your pedagogical strengths and develop your teaching weaknesses.
- 4. From the beginning, practice strictness as a matter of policy and grace as a matter of humanity. Be yourself let students see who you are.
- 5. Create classroom spaces in which everyone feels encouraged to participate. Be willing to learn about and use inclusive teaching practices in order to make belonging a reality.
- 6. Punctuate or inform the journey through course content with "big questions" and "big issues" that grapple with truth and the nature of the absolute.
 - 7. Assign frequent, lower stakes assignments as a way to help students measure their learning progress. Give meaningful feedback on each assignment.
 - 8. Use a midterm course evaluation to garner feedback and improve the course.
 - Be willing to put a lesson plan aside if students really want or need to talk about something, like a campus incident or national event.
 - 10. Remember first, last, and in between that you are teaching people, not the subject. Take every opportunity to show students you care about them as people and about their learning.

Center for Teaching and Learning tips

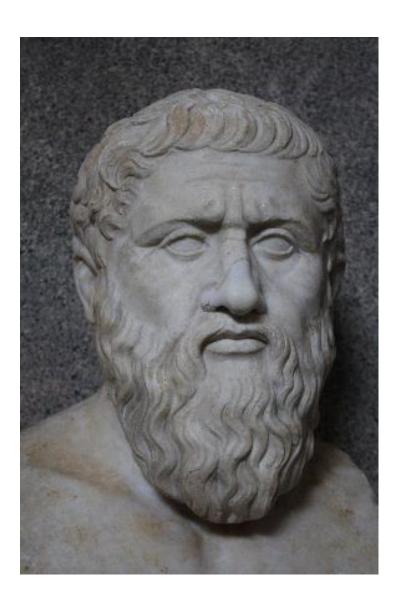
Tip 1: Learn every student's name

Tip 6: Punctuate or inform the journey through the course content with "big questions" and "big issues" that grapple with truth and the nature of the absolute

Quick Review of central concepts in Intro Statistics



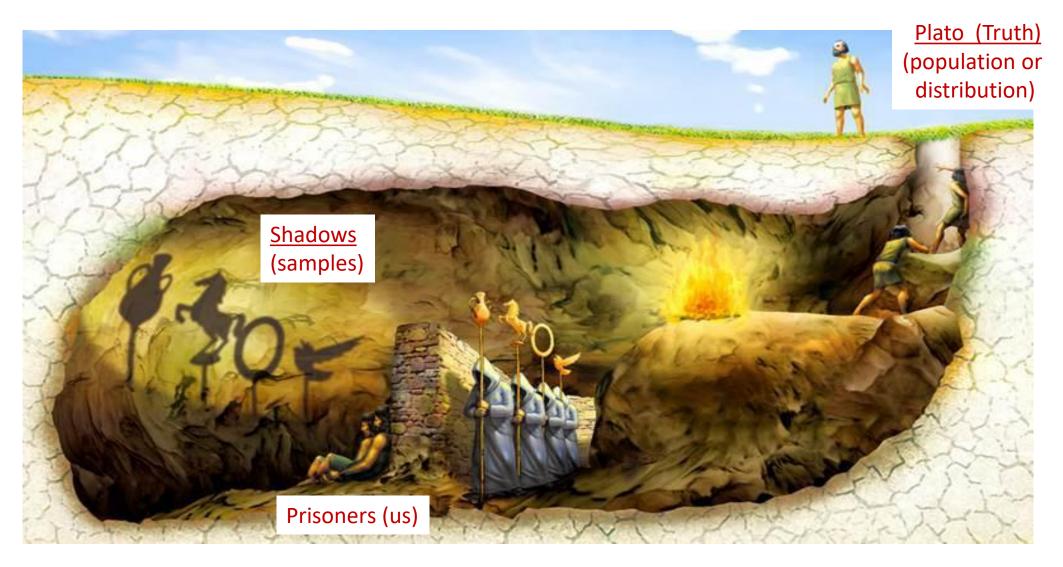
The Truth®!

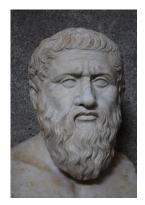


If we could see all the (infinite) data, we would know the Truth®!

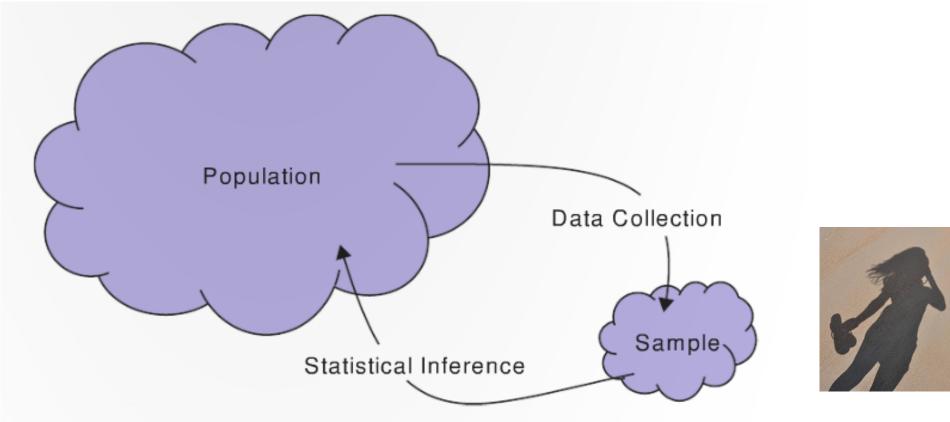
Alas, we can only see a small subset of the data (a sample) so we merely see a shadow of the Truth

Plato's cave

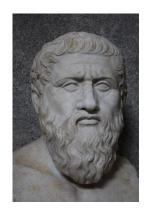




Population: all individuals/objects of interest

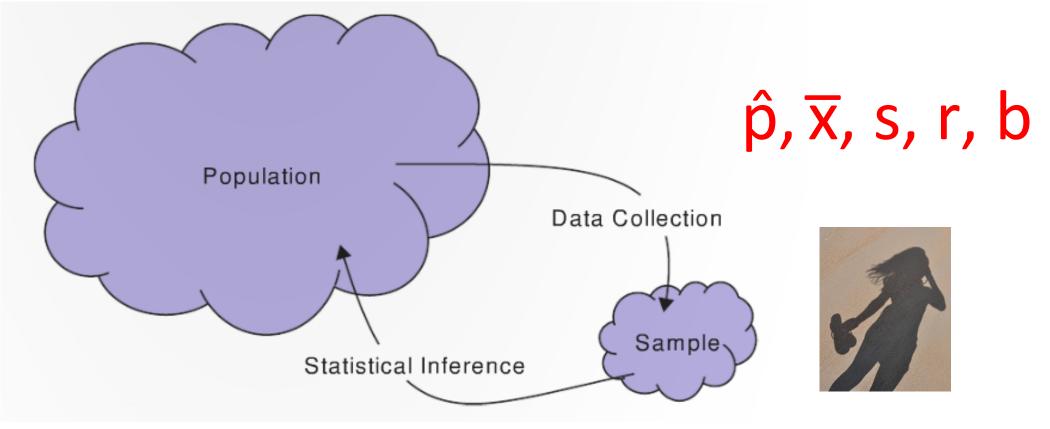


Sample: A subset of the population



π, μ, σ, ρ, β

Parameter: a number characterizing a property of a population



Statistic: A number computed from a sample

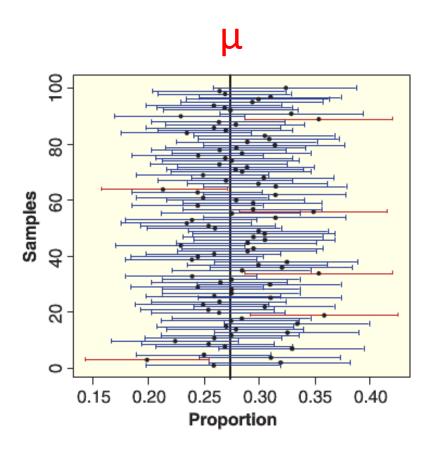
Parameters and statistics commonly used symbols



	Population parameter (Plato)	Sample statistic (shadow)
Mean	μ	x
Standard deviation	σ	S
Proportion	π	ĝ
Correlation	ρ	r
Regression slope	β	b

Inference on parameters

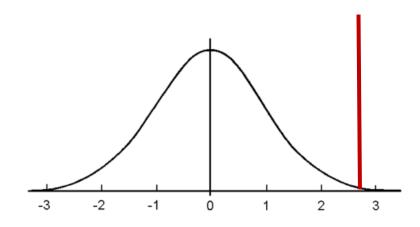
Confidence intervals



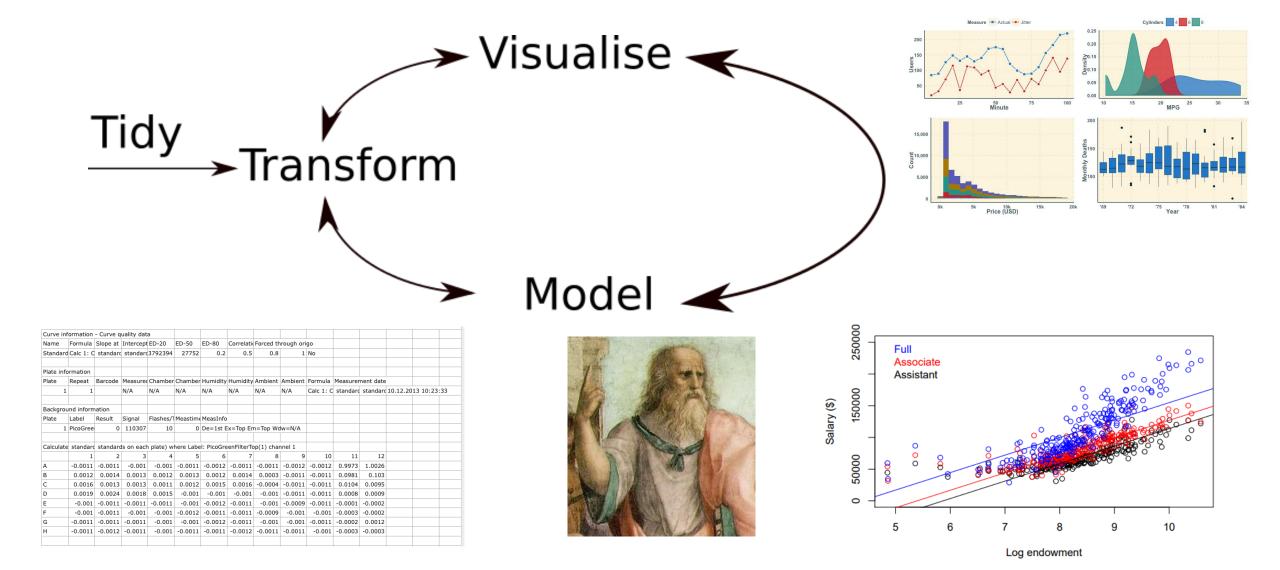
Hypothesis tests

 H_0 : $\mu = 0$

 H_A : $\mu > 0$



Sometimes the Truth is more complicated...





Question



R Basics

Does everyone have R and R Studio installed?

Instructions and a video are on Canvas

Let's take a 2 minute break and open R Studio and follow along...

R and R Studio

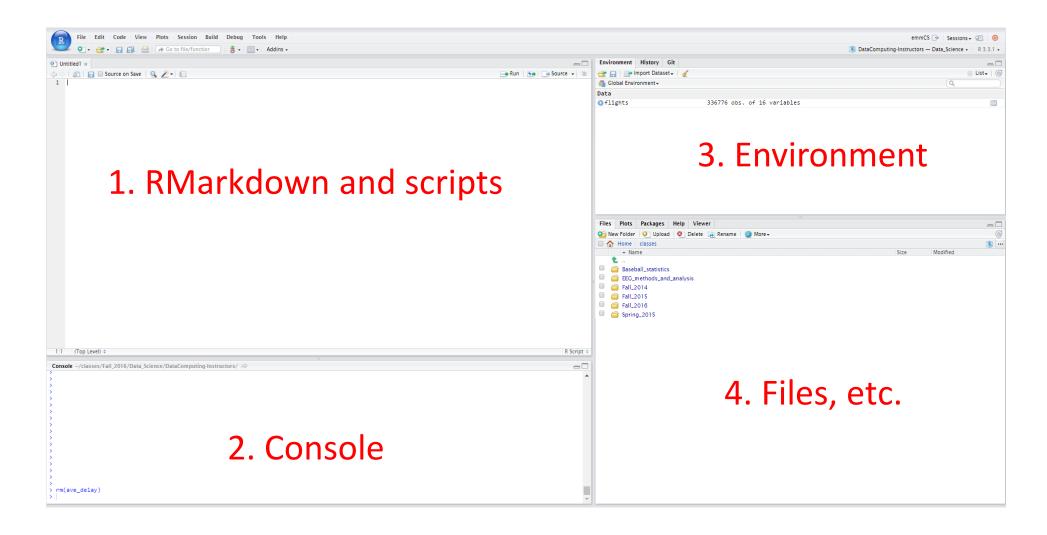
R: Engine



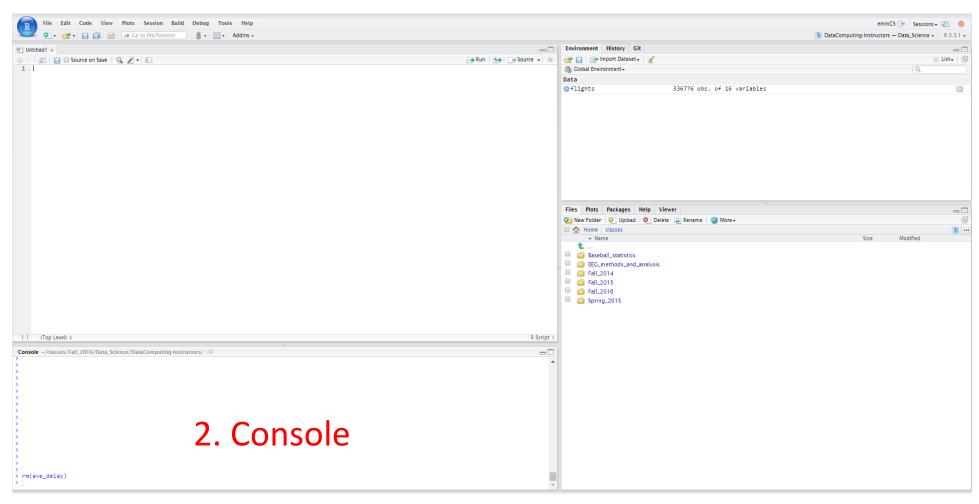
RStudio: Dashboard



RStudio layout



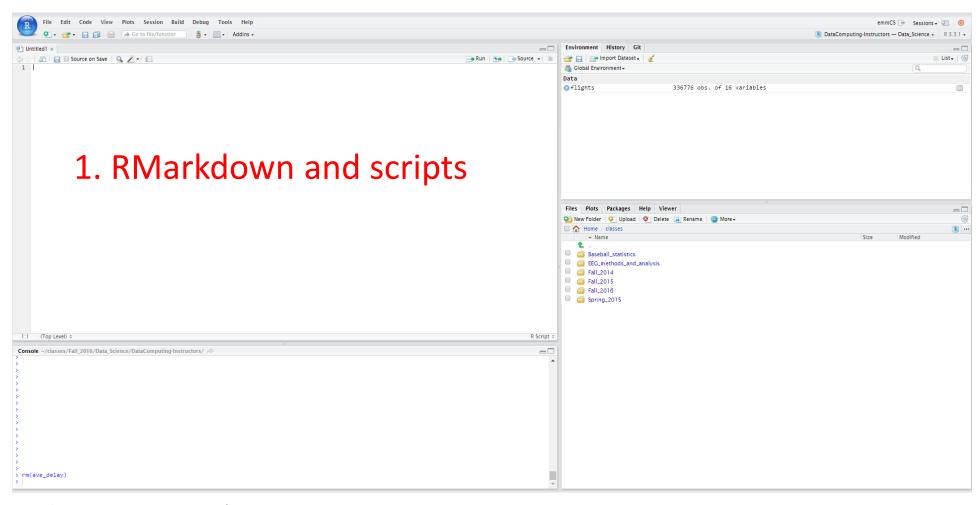
RStudio layout



R as a calculator

- > 2 + 2
- > 7 * 5

RStudio layout



Create a new script

File -> New File -> R Script

Save the script with a reasonable name, e.g., week1_notes.R

R Basics

Arithmetic:

2 + 27 * 5

Assignment of values to *objects*:

> a <- 4
> b <- 7
> z <- a + b
> z
[1] 11

Number journey...

Number journey

```
> a <- 7</li>> b <- 52</li>> d <- a * b</li>> d[1] 364
```

Character strings and Booleans

```
> a <- 7
> s <- "s is a terrible name for an object"
> b <- TRUE
> class(a)
[1] numeric
> class(s)
[1] character
```

Functions

Functions use parenthesis: functionName(x)

```
> sqrt(49)
```

> tolower("DATA is AWESOME!")

To get help

> ? sqrt

One can add comments to your code

> sqrt(49) # this takes the square root of 49

Vectors

Vectors are ordered sequences of numbers or letters The c() function is used to create vectors

```
> v <- c(5, 232, 5, 543)
> s <- c("statistics", "data", "science", "fun")
```

One can access elements of a vector using square brackets [] > s[4] # what will the answer be?

We can get multiple elements from a vector too > s[c(1, 2)]

Vectors continued

One can assign a sequence of numbers to a vector

- > z <- 2:10
- > z[3]

One can test which elements are greater than a value

Can add names to vector elements

```
> names(v) <- c("first", "second", "third", "fourth")
```

Vectors continued

One can also apply functions to vectors

- > z <- 2:10
- > sqrt(z)
- > mean(z)

Questions?



R packages

Packages add additional functionality to R



We will use many additional packages in this class

• gplyr, ggplot2, tidyr, etc.

There is also a class specific package (SDS230) I wrote that you can use to download homework and other files

All class materials are also on GitHub: https://github.com/emeyers/SDS230

Installing SDS230 package and LaTeX

To install the SDS230 package you first need to install the devtools package which can be done using:

install.packages("devtools")

You can then install the class SDS230 package using the function:

devtools::install_github("emeyers/SDS230")

Installing SDS230 package and LaTeX

Finally, after you have installed the SDS package, there is a function in the SDS package that installs LaTeX on you computer

(this function uses the tinytex package)

To install LaTeX use:

```
SDS230:::initial_setup() # will packages you need for the class tinytex:::install_tinytex() # will install LaTeX via tinytex package
```

Test that the installation worked

```
tinytex:::is_tinytex() # will return TRUE if it works (note: 3 colons)
```

For next class

- 1. If you have not done so already
 - Fill out class survey on Canvas under the Quizzes link
 - Install R and RStudio if you have not done so already

2. Install the SDS230 class package and LaTeX

Questions?

