
Data
Exploration
and
Analysis

Overview

Introductions

Overview and logistics of the course

Review of a few central concepts from Intro Stats

Introduction to R

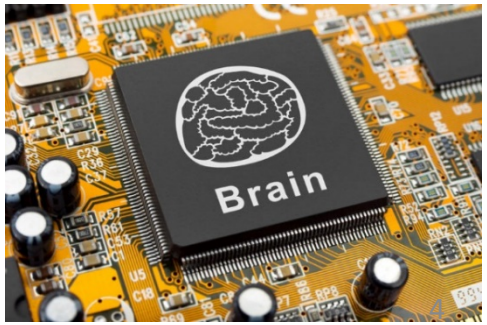
- R as a calculator
- Objects and vectors
- Installing the class SDS230 package and LaTeX


TIME
FOR
INTRODUCTIONS

Ethan Meyers (he/him)

- Visiting Associate Professor at Yale
- Associate Professor of Statistics Hampshire College
- Research Affiliate at the Center for Brains, Minds and Machines at MIT

Research area: Machine learning to analyze neural data



Ethan.Meyers@yale.edu

Teaching Assistants

Teaching Fellows (TF)

- Hayon Michelle Choi: hayonmichelle.choi@yale.edu
- Megan Ayers: m.ayers@yale.edu

Undergraduate Learning Assistants (ULA)

- Marc-Henry: marc-henry.dorval@yale.edu
- Nathan Kim: nathan.kim@yale.edu
- Jalen Li: jalen.li@yale.edu
- Stephan Billingslea: stephan.billingslea@yale.edu
- Mercy Idindili: mercy.idindili@yale.edu
- Stephanie Hu: stephanie.hu@yale.edu
- Gemma Yoo: gemma.yoo@yale.edu
- Rebecca Del Rio: rebecca.delrio@yale.edu


What is this class about?

Course objectives

Extend and solidify concepts and method learned in intro stats

- Non-parametric tests, multiple regression, etc.
- Focus on insights and why methods work rather than proofs

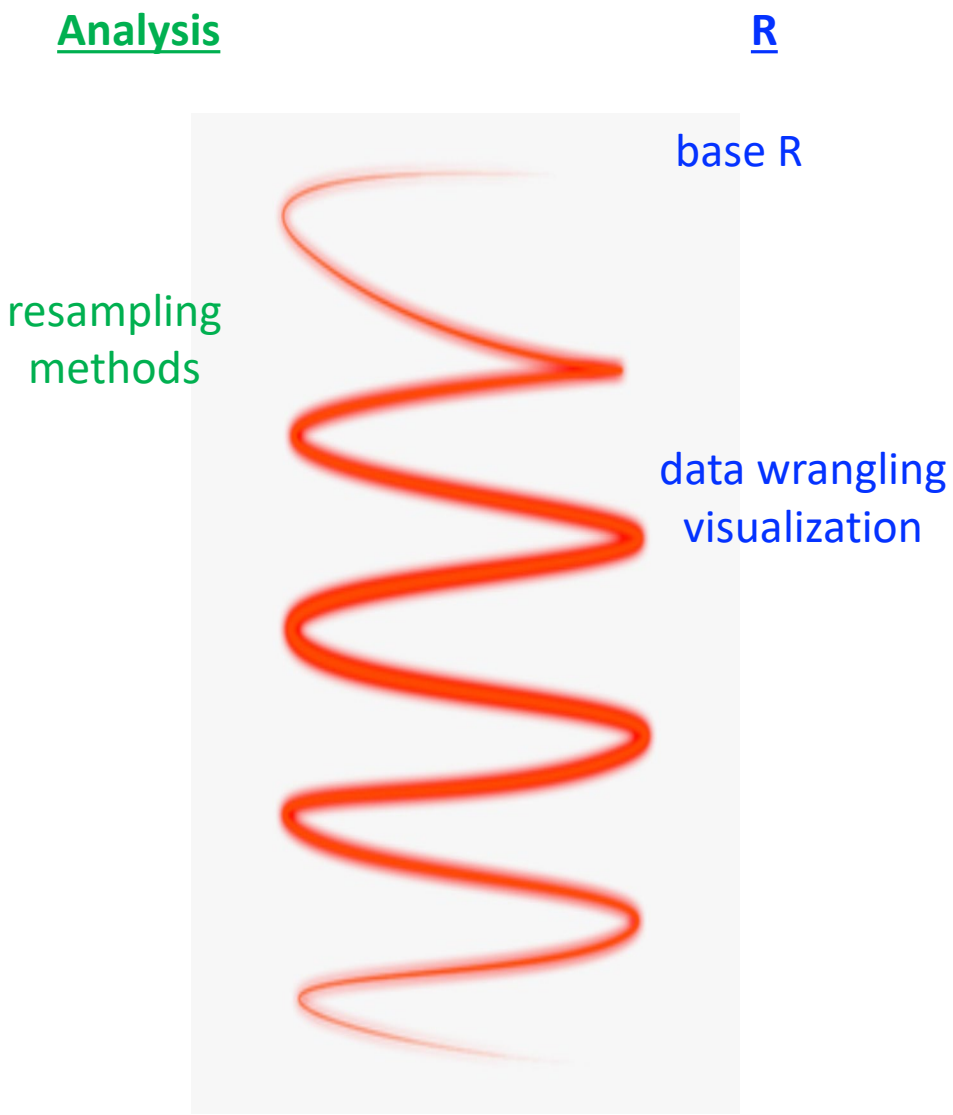
Learn how to use the R programming language to analyze, visualize and wrangle data

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!

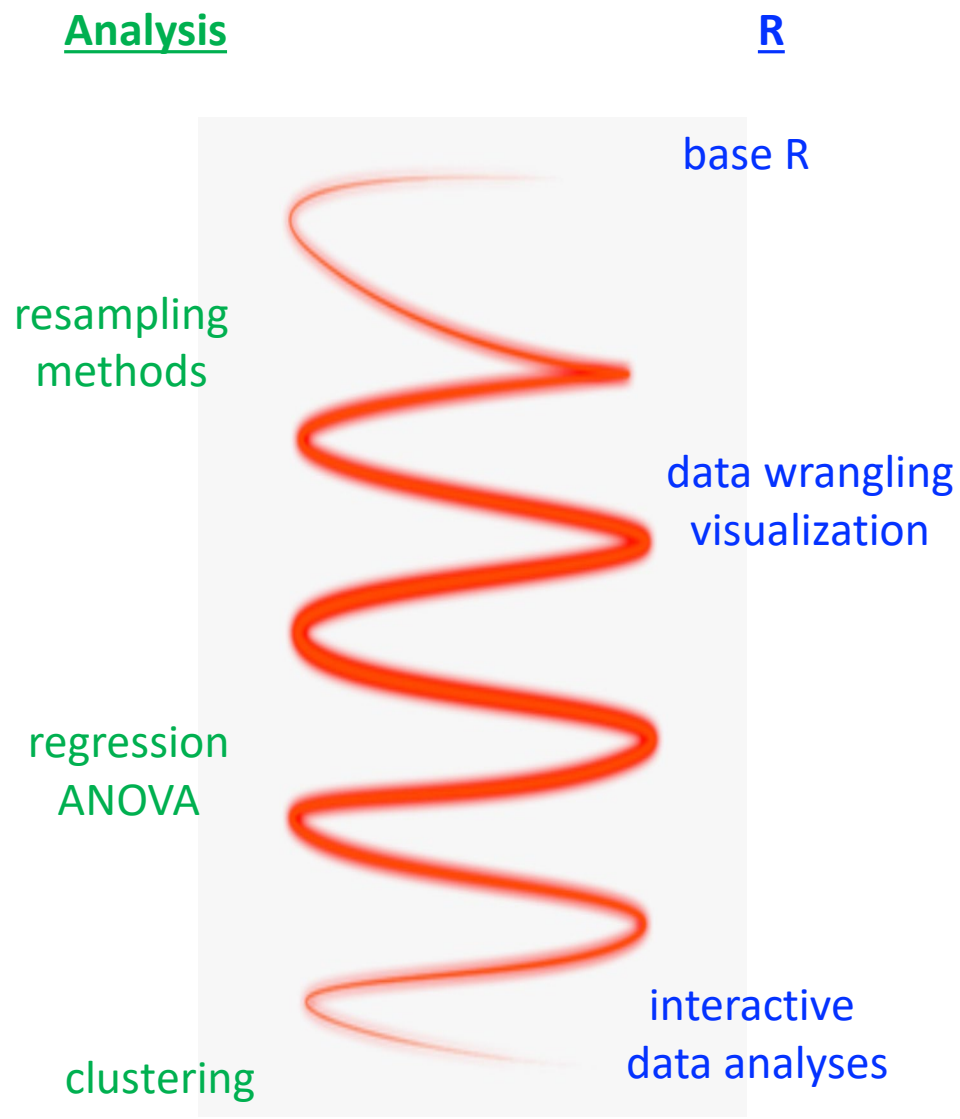
Plan for the semester

1	Sep 2	Course overview, introduction to R, descriptive statistics
2	Sep 7-9	Review of central statistical concepts and exploratory analysis using R
3	Sep 14-16	Confidence Intervals and the bootstrap
4	Sep 21-23	Review of hypothesis tests and permutation tests in R
5	Sep 28-30	Parametric, non-parametric and theories of hypothesis testing
6	Oct 5-7	Data manipulation and visualization
7	Oct 12-14	Mapping, review and midterm exam
8	Oct 19-22	October break



Plan for the semester

9	Oct 26-28	Simple linear regression
10	Nov 2-4	Multiple regression
11	Nov 9-11	Analysis of Variance
12	Nov 16-18	Data cleaning (I'm away at a conference) and logistic regression
13	Nov 22-26	November break
14	Nov 30 Dec 2	Clustering and principal components analysis
15	Dec 7-9	Ethics, interactive data analyses, wrap up and review
16	Dec 21st	Final exam at 7pm



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, logistic regression

ANOVA: one-way/factorial, interactions

Statistical learning: cross-validation, logistic regression, PCA, clustering

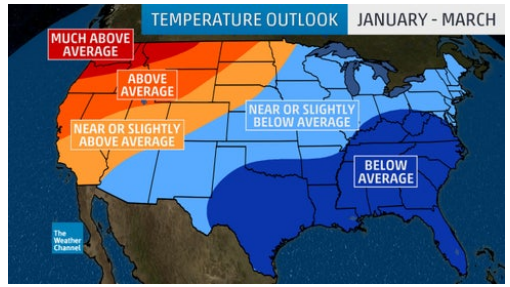
Examples of questions we might look at…

Bootstrap confidence intervals: How much do avocados typically cost?

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

Data summarization: which airlines have the longest flight delays?

Data wrangling/visualization: How accurate are weather predictions?



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

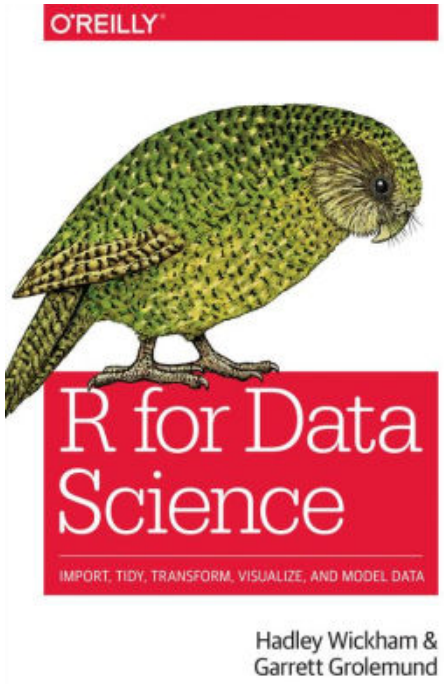
Class structure

Class time 9-10:15am Tuesdays and Thursdays

- Will introduce new content during class
- Class activities, reading discussions, question answering, etc.
- Might have some pre-recorded video content to be watched prior to class

Canvas website:
https://yale.instructure.com/courses/68751

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



Office hours

My planned office hours (subject to change)

- Office hours will be on zoom and in 24 Hillhouse room 206

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 specific questions about content in the class, best to first ask them on Ed Discussion

- Class participation grade based on questions and answers on Ed Discussion

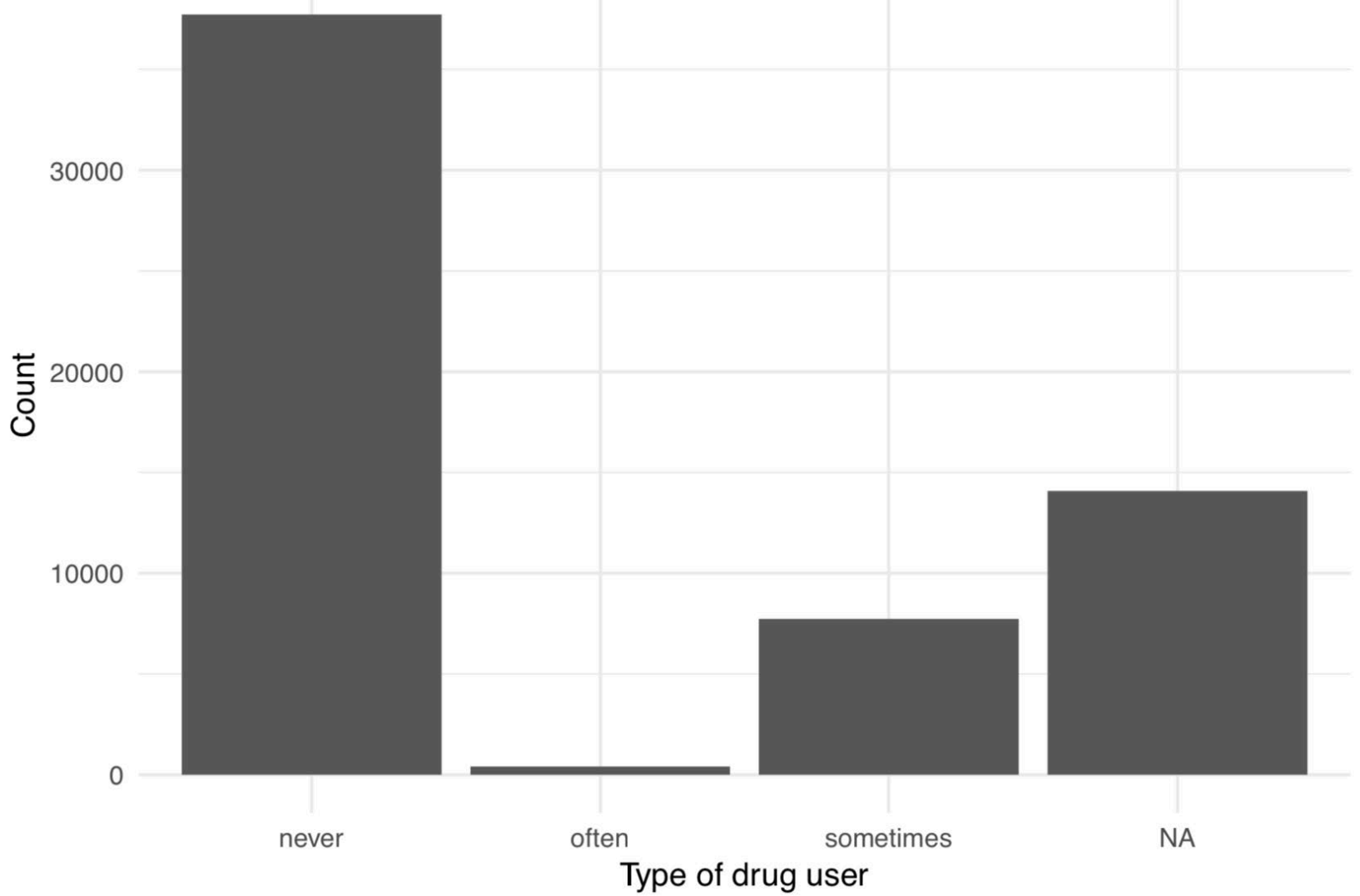
Assignments and grades

1. Homework problem sets (54%)
   - Exploring concepts and analyzing data using R
   - Weekly: 10 total

Homework policies

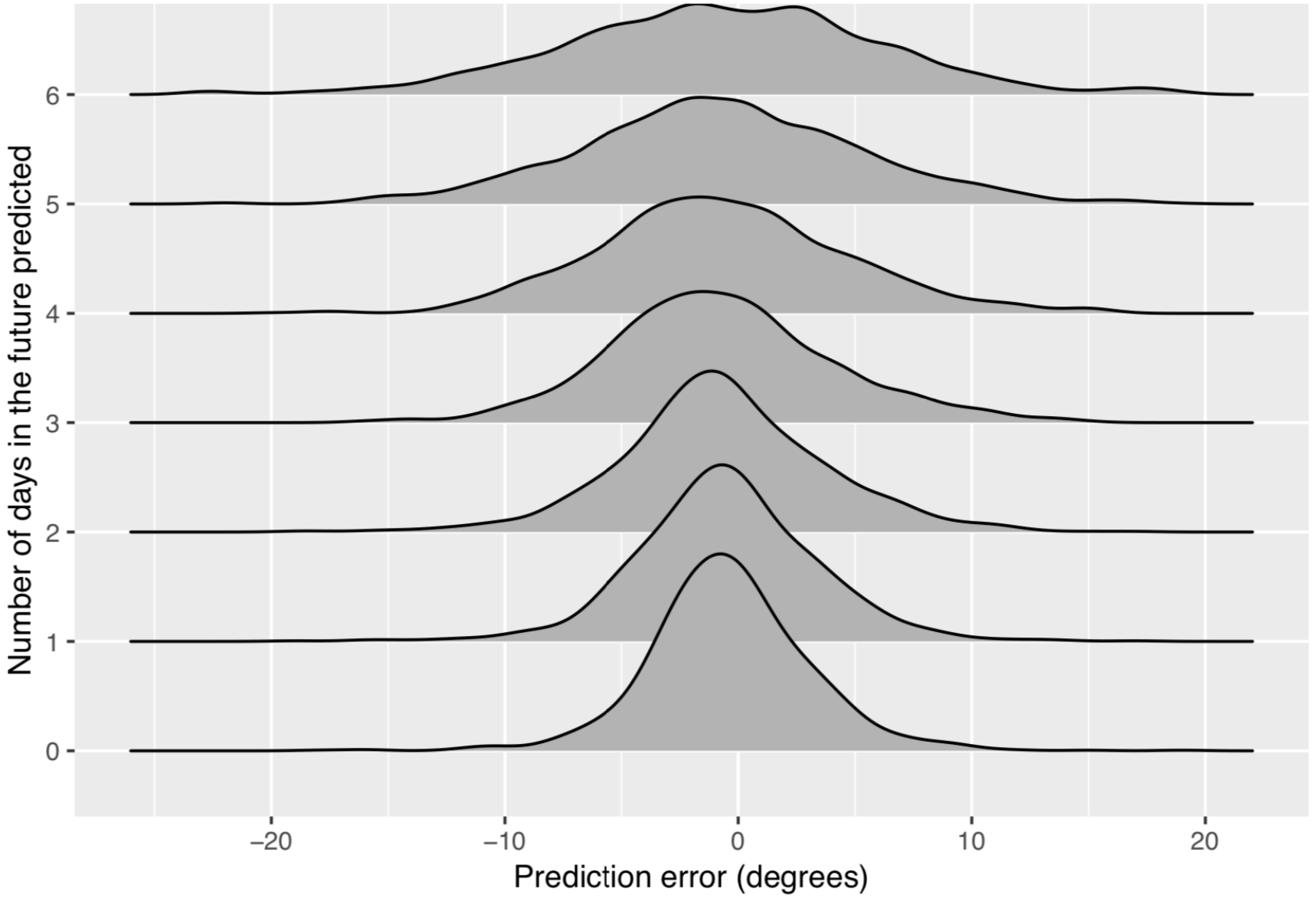
- You may discuss questions with other but the work you turn in must be your own
- Homework assigned on Tuesdays and are due at 11pm on Sundays
- Late worksheets (90%) credit if turned in by 11:59pm on Monday
  - For any other extension a Dean's Excuse is needed
- Lowest scoring worksheet will be dropped

Typical homework assignment piece




```
# Bonus: create a pie chart of the self reported frequency of
# drug use and make it look good!
profiles %>% count(drugs) %>% filter(!is.na(drugs)) %>% ggplot(aes(x = "",
    y = n, fill = drugs)) + geom_col(width = 1) + coord_polar(theta = "y") +
    theme_minimal() + theme(axis.title.x = element_blank(), axis.text.x = element_blank(),
    axis.ticks.x = element_blank(), panel.grid.major = element_blank(),
    panel.grid.minor = element_blank()) + xlab("")
```

Typical 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 (8%)
   - Find a data set and analyze it on your own (5-7 page report)

3. Exams (35% total)
   - Midterm (15%) Oct 14th during class
   - Final (20%) Dec 21st at 7pm

4. Participation (3%)
   - Active asking and answering questions on Ed Discussions
   - Engaging with short class readings and discussions

Class survey

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

- Under the Quizzes link on the left on Canvas

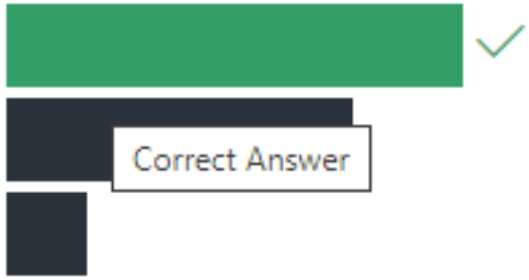
Preliminary class survey results

As of 7pm yesterday, 67 people had filled out the class survey

- ~67% of the class is undergraduates,
- ~30% graduate students

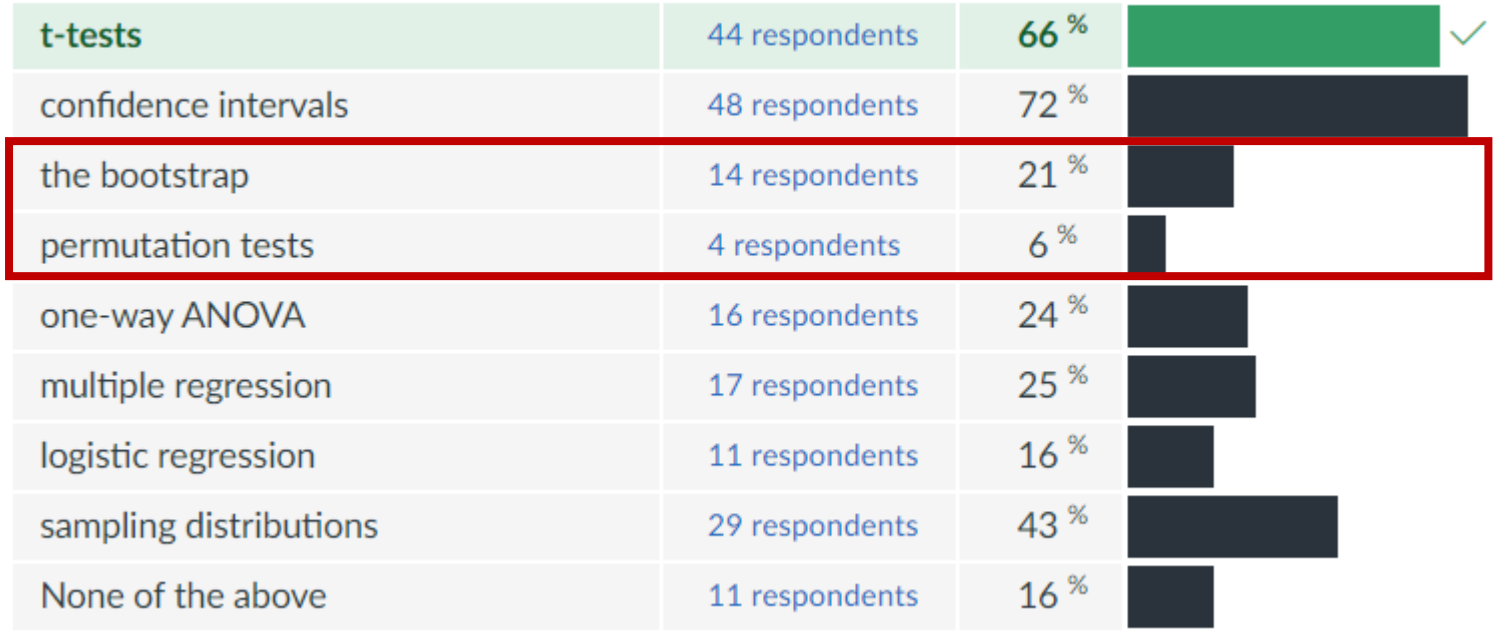
Have you taken an introductory Statistics class before?

Yes, in college	40 respondents	**60 %**	✓
Yes, in high school (AP stats)	30 respondents	45 %	
No	6 respondents	9 %	



Class survey results

Which Statistics methods/concepts are you comfortable with?



Class survey results

Have you taken linear algebra and/or calculus?

- These are not prerequisites
- Programming experience is also not a prerequisite

Yes, linear algebra	11 respondents	**16 %**	
Yes, calculus I/II	31 respondents	46 %	
Yes, multivariable calculus	18 respondents	27 %	
Yes, I have taken probability	3 respondents	4 %	
I have not taken any of these classes	4 respondents	6 %	

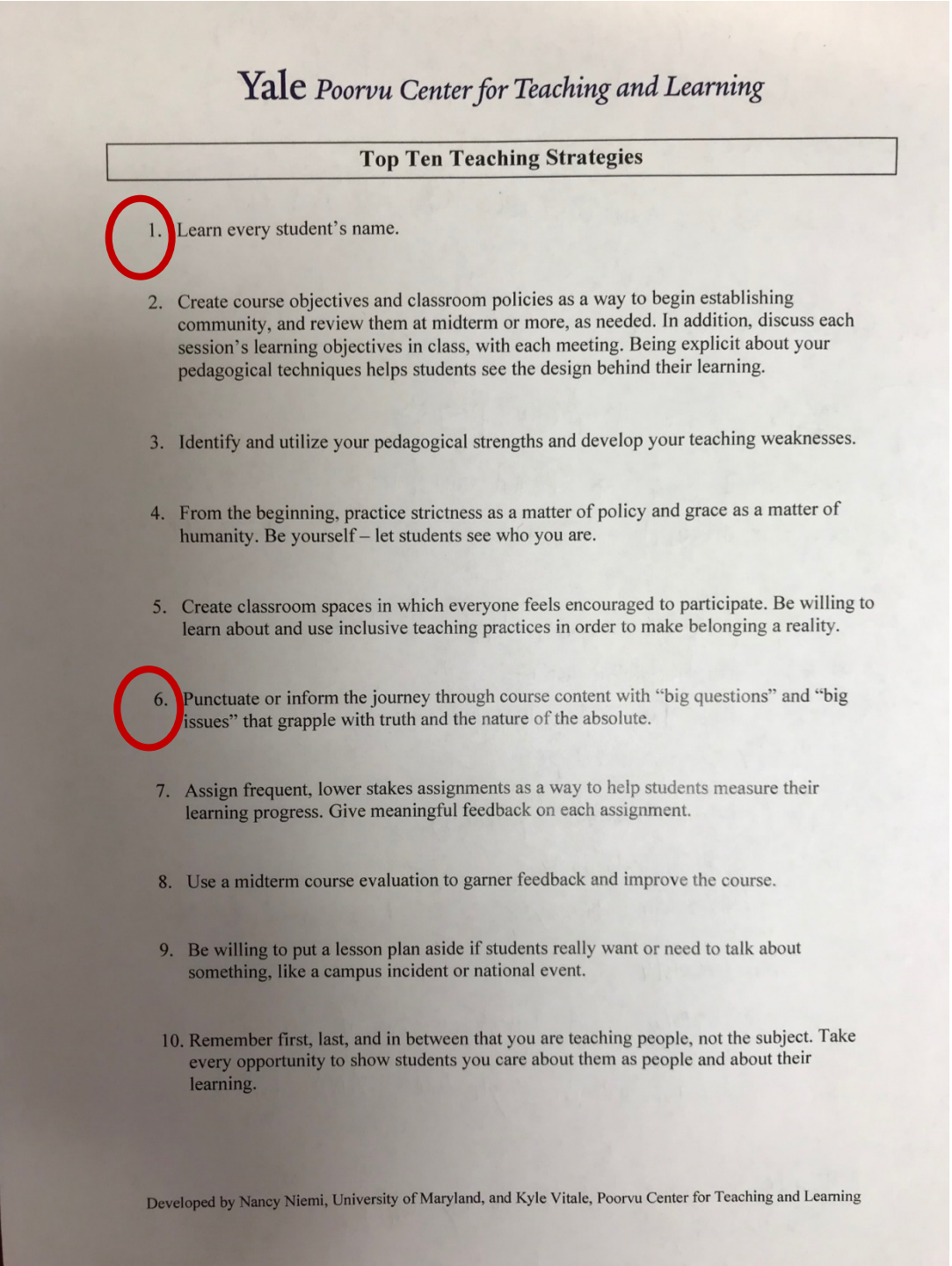
Class survey results

How much experience do you have with computer programming?

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


ANY QUESTIONS?

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.
9. 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.

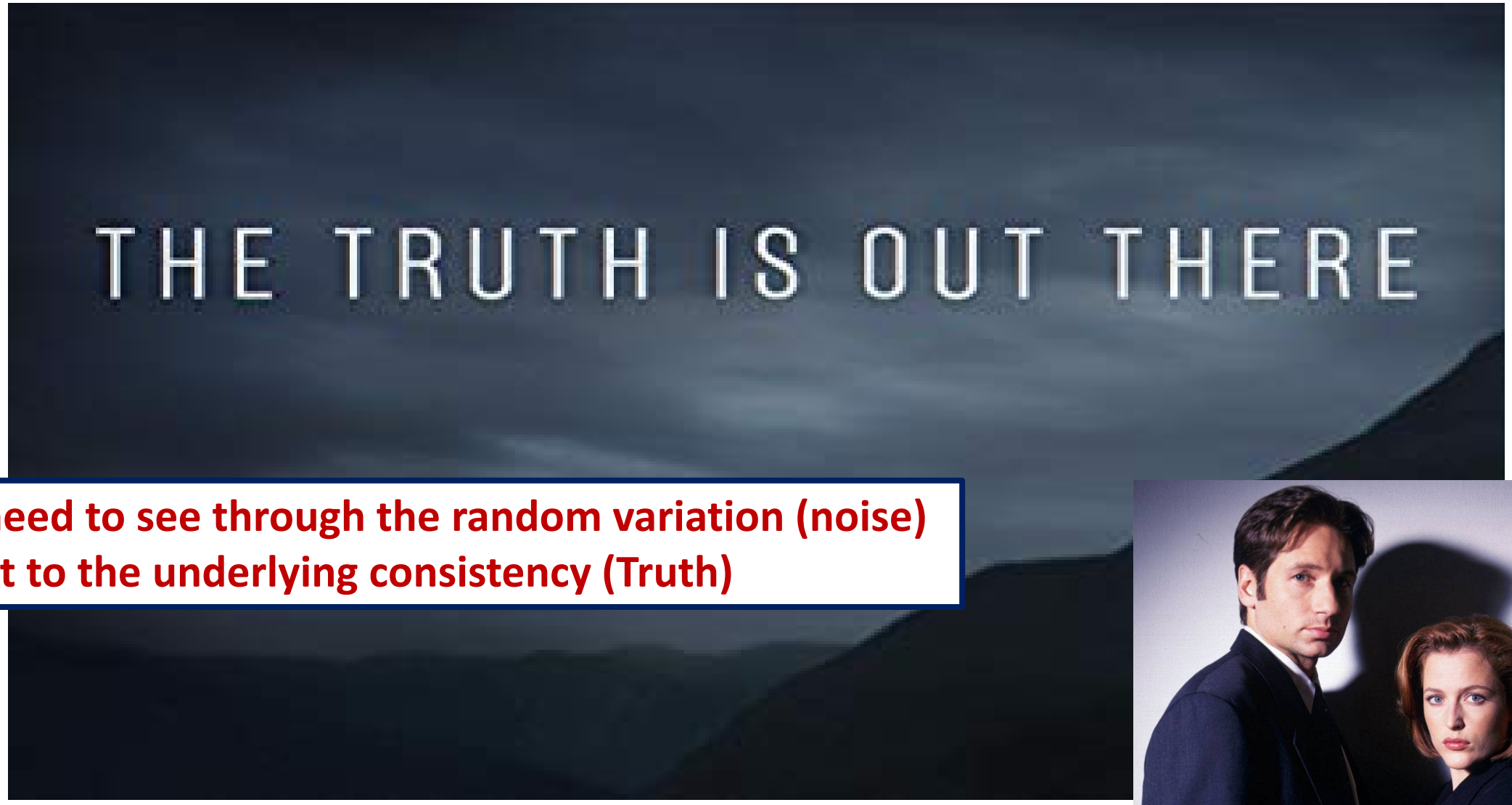
Developed by Nancy Niemi, University of Maryland, and Kyle Vitale, Poorvu Center for Teaching and Learning

Center for Teaching and Learning tips

Tip 1: Learn every student's name

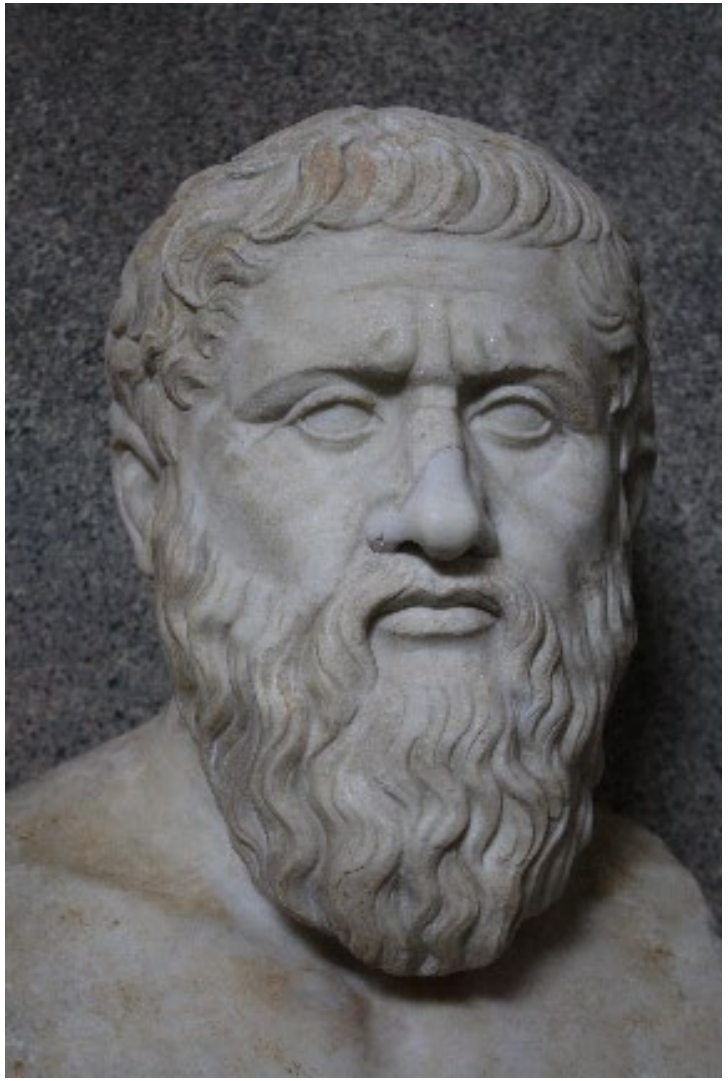
Tip 6: Punctuate or inform the journey through the course content with "big questions" and "big issues" that grapple with truth and the nature of the absolute

Quick Review of central concepts in Intro Statistics



We need to see through the random variation (noise) to get to the underlying consistency (Truth)

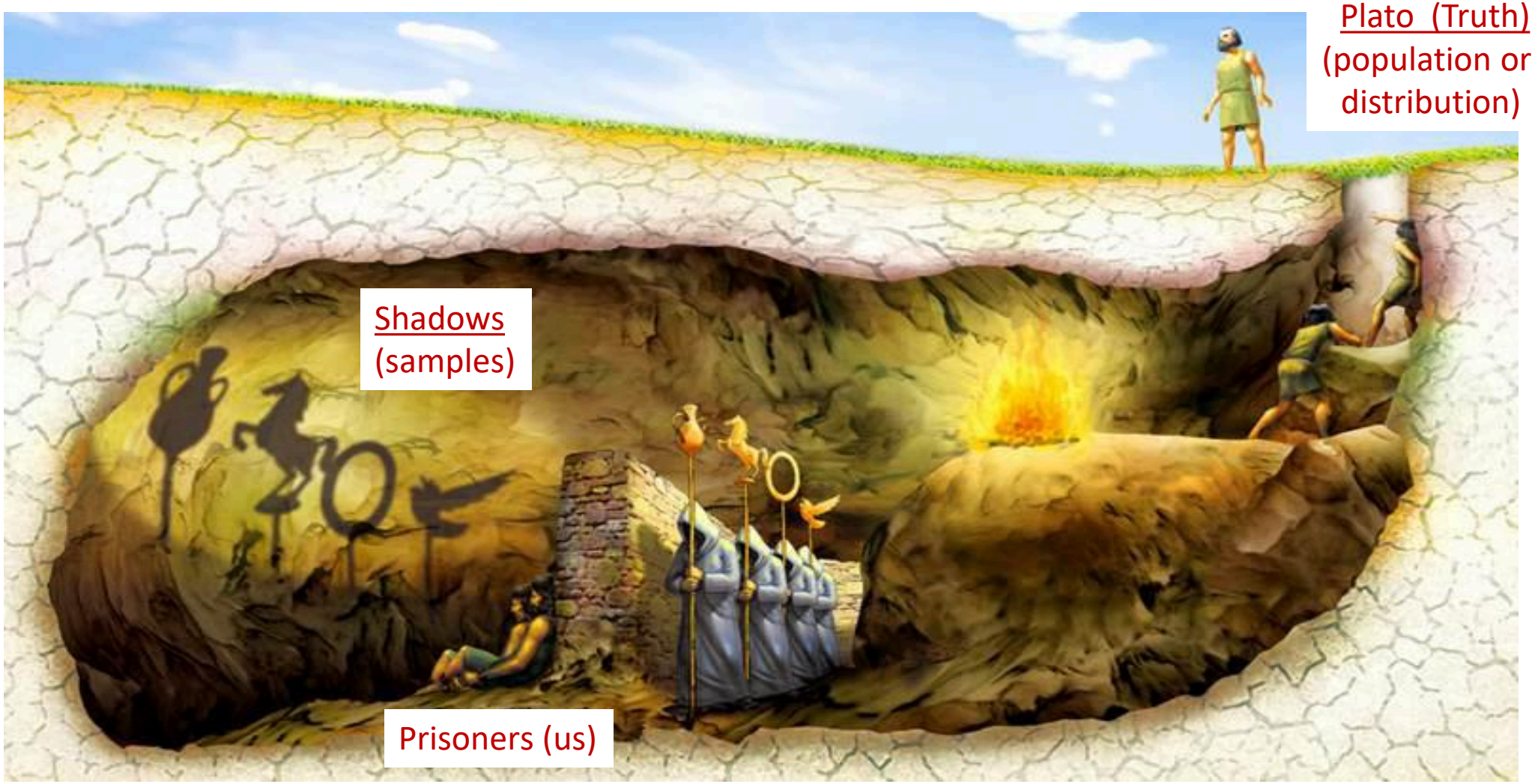
The Truth®!



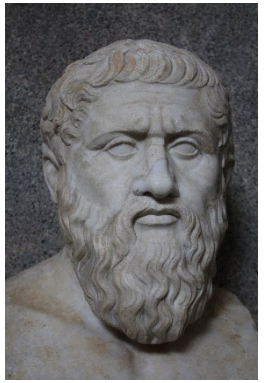
If we could see all the (infinite) data, we would know the Truth®!

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

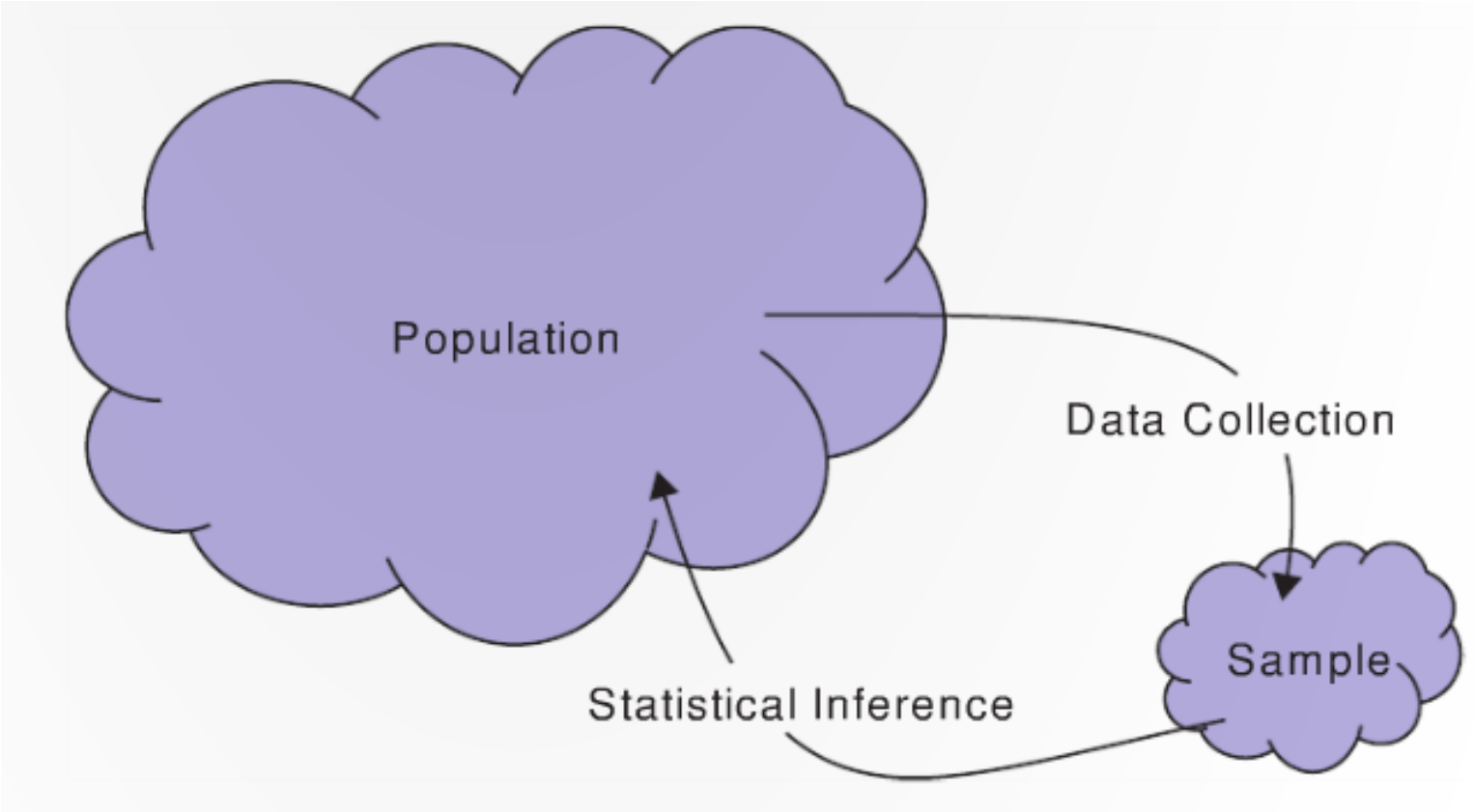
Plato's cave

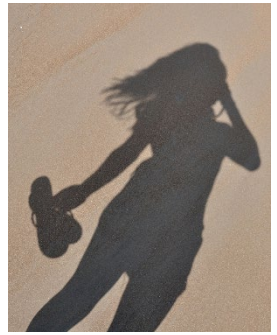


From The Republic (~ 380 BCE)

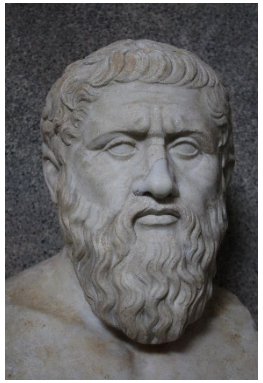


Population: all individuals/objects of interest



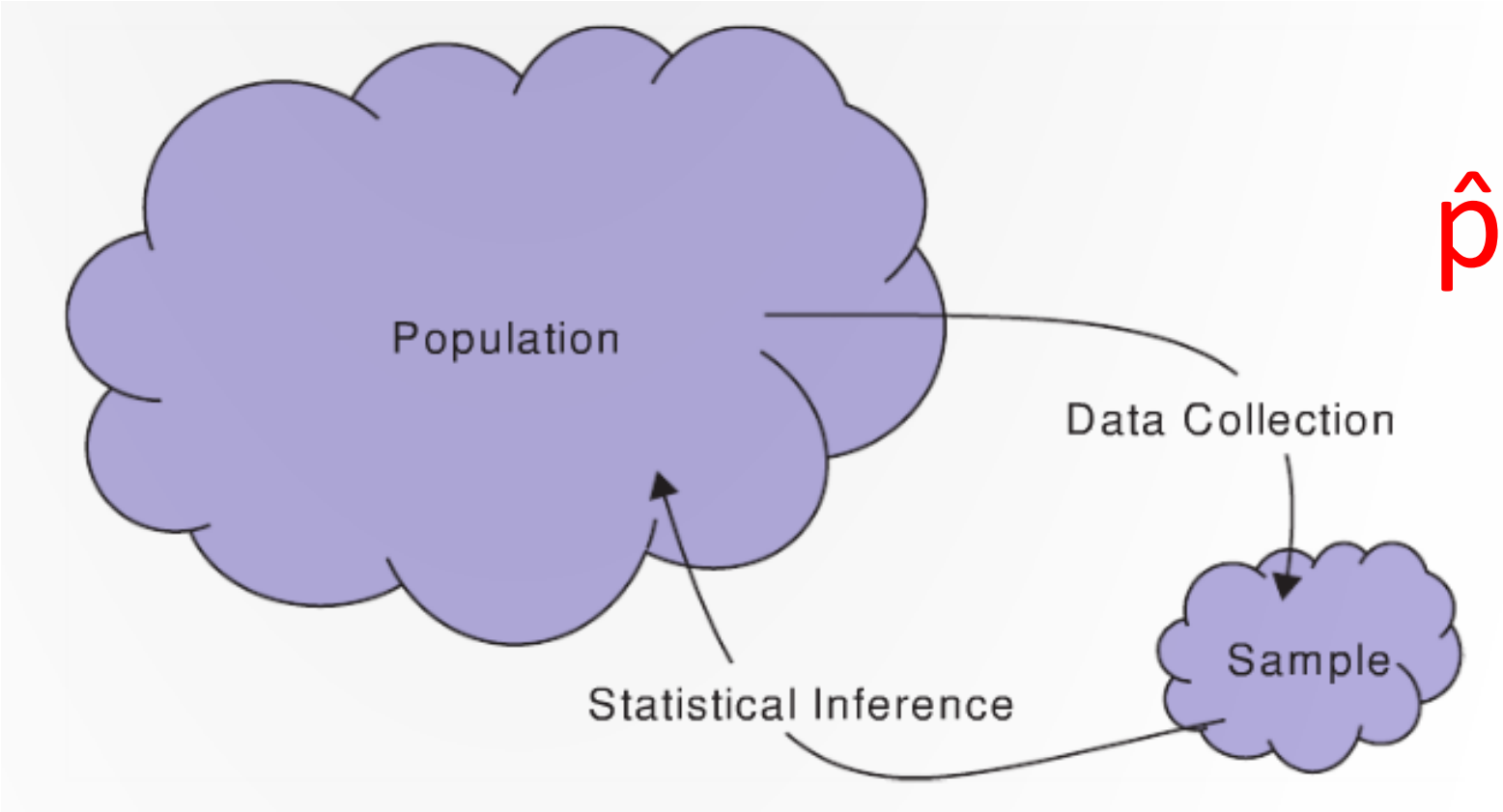


Sample: A subset of the population

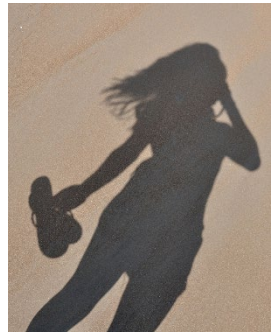


$\pi, \mu, \sigma, \rho, \beta$

Parameter: a number characterizing a property of a population

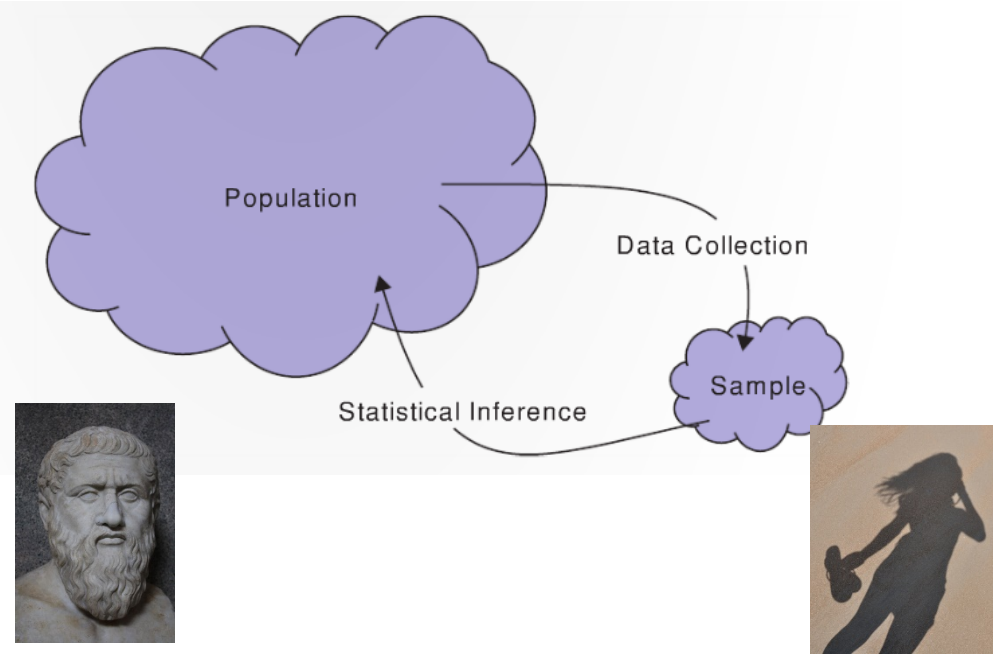


$\hat{p}, \bar{x}, s, r, b$



Statistic: A number computed from a sample

Parameters and statistics commonly used symbols

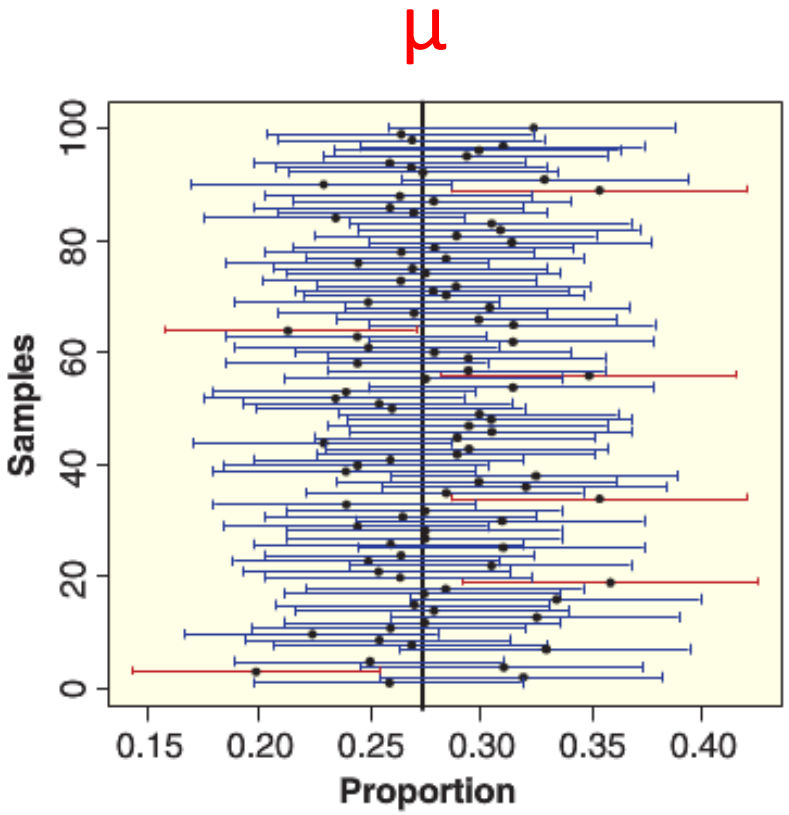


$$\bar{x} = \frac{\sum_i^n x_i}{n}$$

	Population parameter (Plato)	**Sample statistic (shadow)**
Mean	μ	x̄
Standard deviation	σ	s
Proportion	π	p̂
Correlation	ρ	r
Regression slope	β	b

Inference on parameters

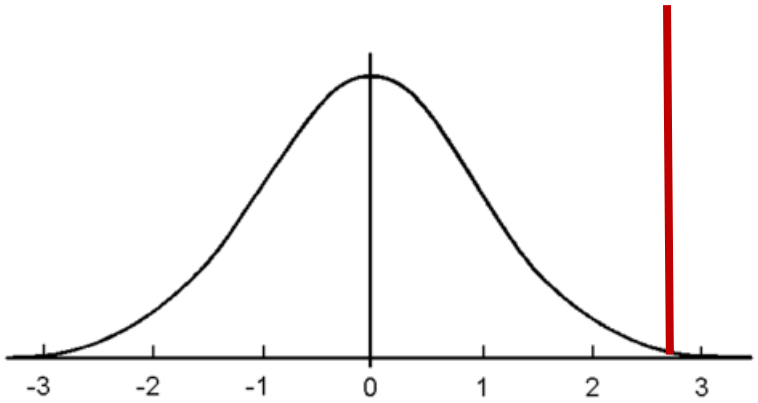
Confidence intervals



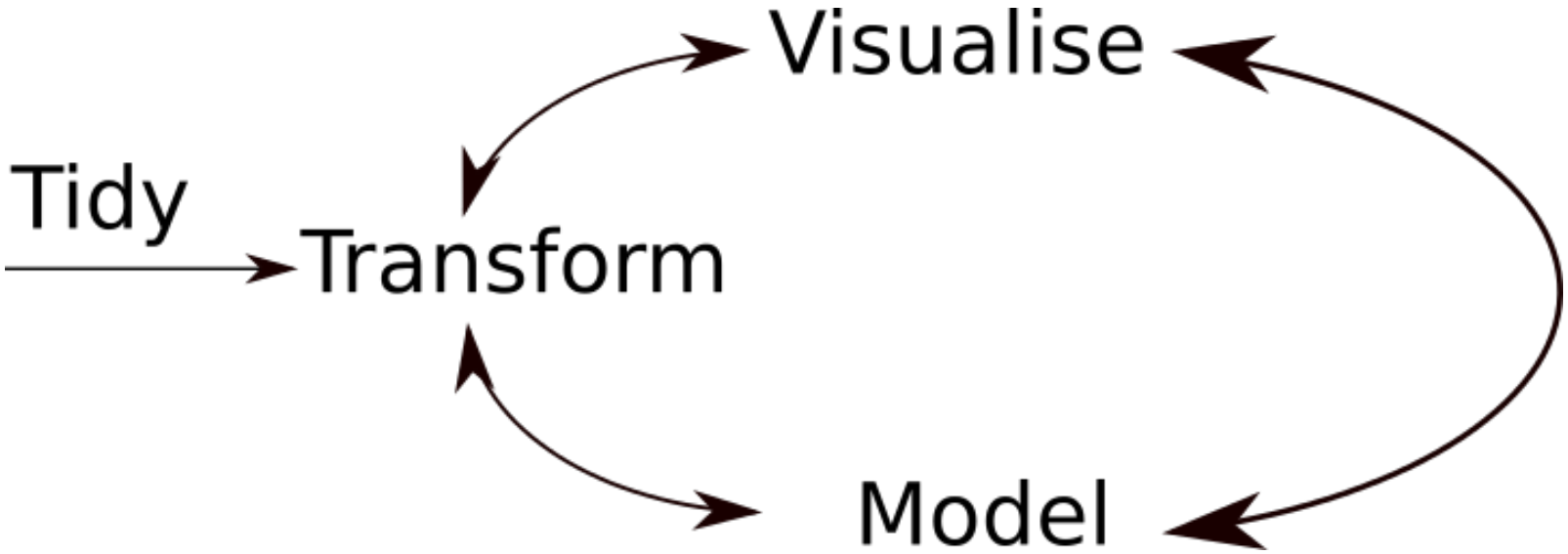
Hypothesis tests

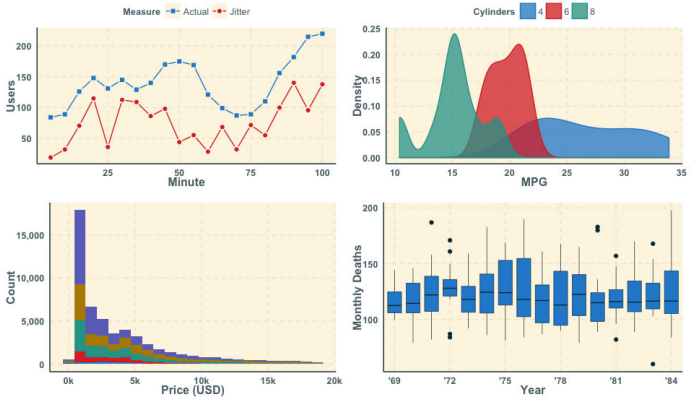
H_0: $\mu = 0$

H_A: $\mu > 0$

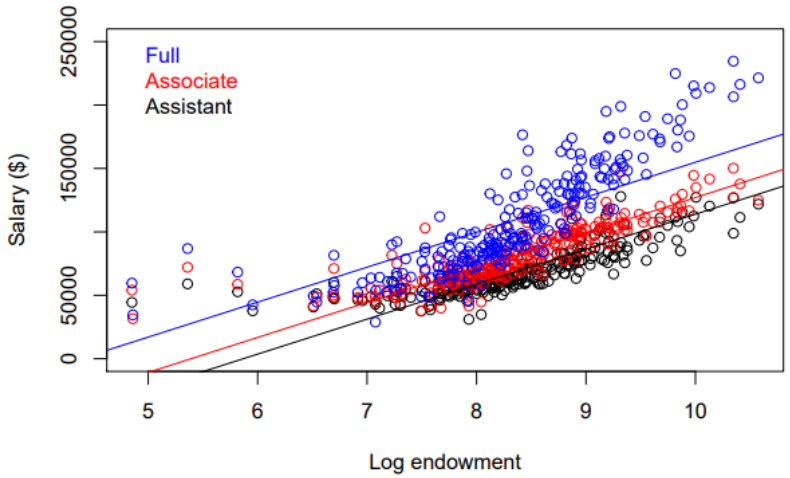


Sometimes the Truth is more complicated…





Curve information - Curve quality data												
Name	Formula	Slope at	Intercept	ED-20	ED-50	ED-80	Correlatic	Forced through origo				
Standard	Calc 1: C	standarc	standarc	3792394	27752	0.2	0.5	0.8	1	No		
Plate information												
Plate	Repeat	Barcode	Measurec	Chamber	Chamber	Humidity	Humidity	Ambient	Ambient	Formula	Measurement date	
1	1		N/A	N/A	N/A	N/A	N/A	N/A	N/A	Calc 1: C	standarc	standarc 10.12.2013 10:23:33
Background information												
Plate	Label	Result	Signal	Flashes/1	Measime	MeasInfo						
1	PicoGree	0	110307	10	0	De=1st Ex=Top Em=Top Wdw=N/A						
Calculate	standarc	standards on each plate) where Label: PicoGreenFilterTop(1) channel 1										
	1	2	3	4	5	6	7	8	9	10	11	12
A	-0.0011	-0.0011	-0.001	-0.001	-0.0011	-0.0012	-0.0011	-0.0011	-0.0012	-0.0012	0.9973	1.0026
B	0.0012	0.0014	0.0013	0.0012	0.0013	0.0012	0.0014	0.0003	-0.0011	-0.0011	0.0981	0.103
C	0.0016	0.0013	0.0013	0.0011	0.0012	0.0015	0.0016	-0.0004	-0.0011	-0.0011	0.0104	0.0095
D	0.0019	0.0024	0.0018	0.0015	-0.001	-0.001	-0.001	-0.001	-0.0011	-0.0011	0.0008	0.0009
E	-0.001	-0.0011	-0.0011	-0.0011	-0.001	-0.0012	-0.0011	-0.001	-0.0009	-0.0011	-0.0001	-0.0002
F	-0.001	-0.0011	-0.001	-0.001	-0.0012	-0.0011	-0.0011	-0.0009	-0.001	-0.001	-0.0003	-0.0002
G	-0.0011	-0.0011	-0.0011	-0.001	-0.001	-0.0012	-0.0011	-0.001	-0.001	-0.0011	-0.0002	0.0012
H	-0.0011	-0.0012	-0.0011	-0.001	-0.0011	-0.0011	-0.0012	-0.0011	-0.0011	-0.001	-0.0003	-0.0003




Questions?

Question

Q: What programming language do pirates use?

R Basics

Does everyone have R and R Studio installed?

- Instructions and a video are on Canvas

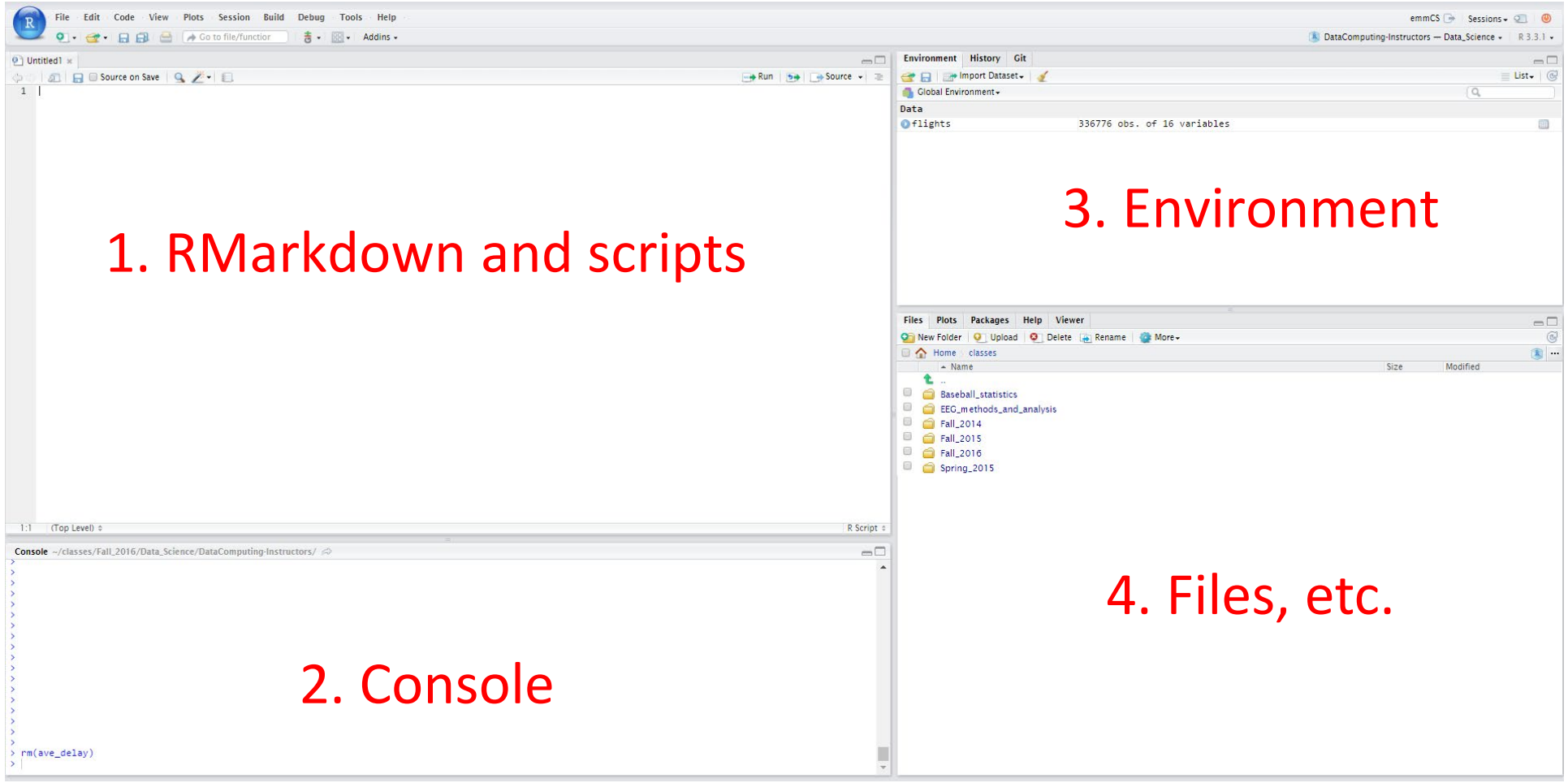
Let’s take a 5 minute break, and then open R Studio and follow along…

R and R Studio

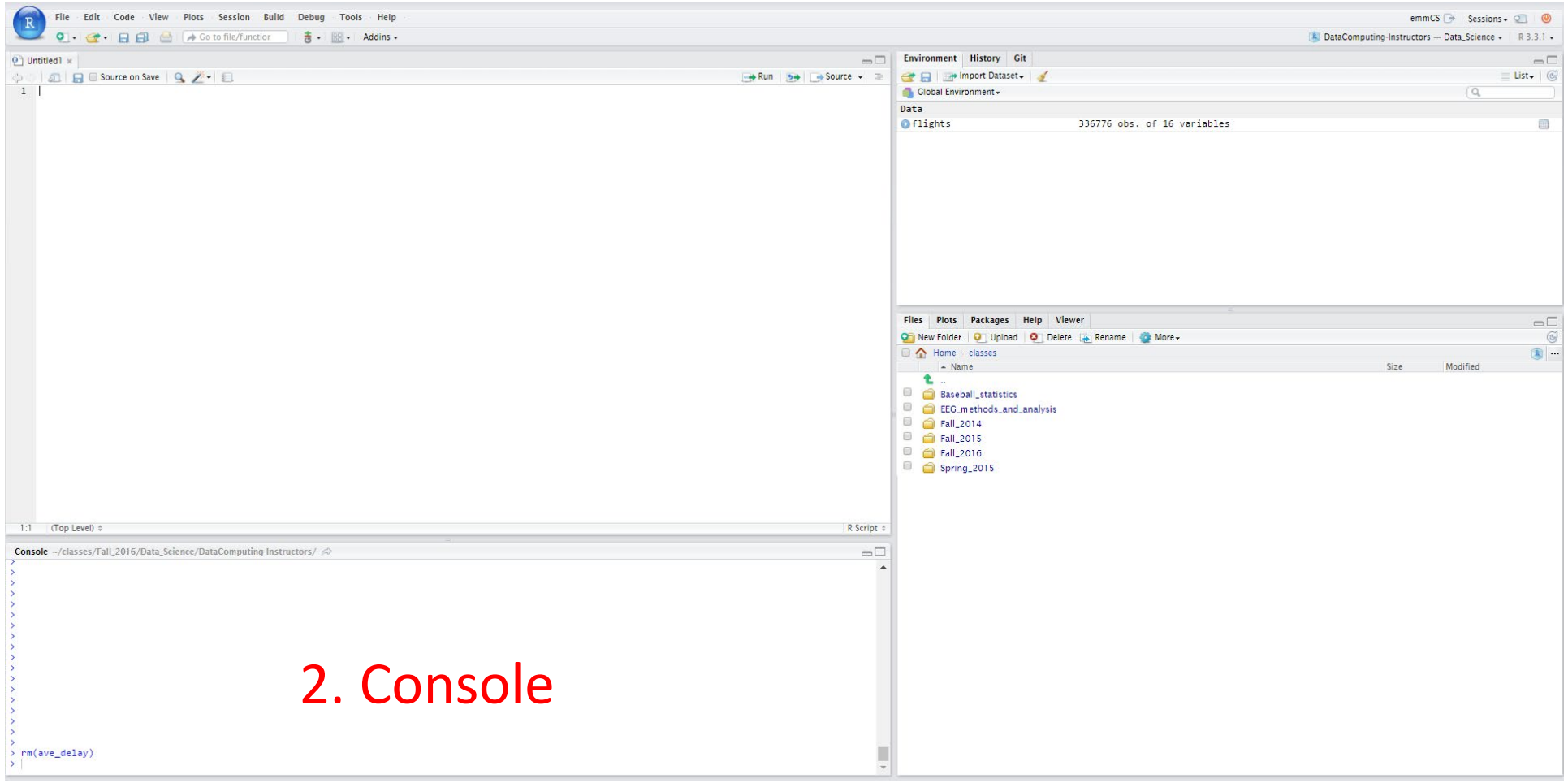
R: Engine

RStudio: Dashboard

RStudio layout



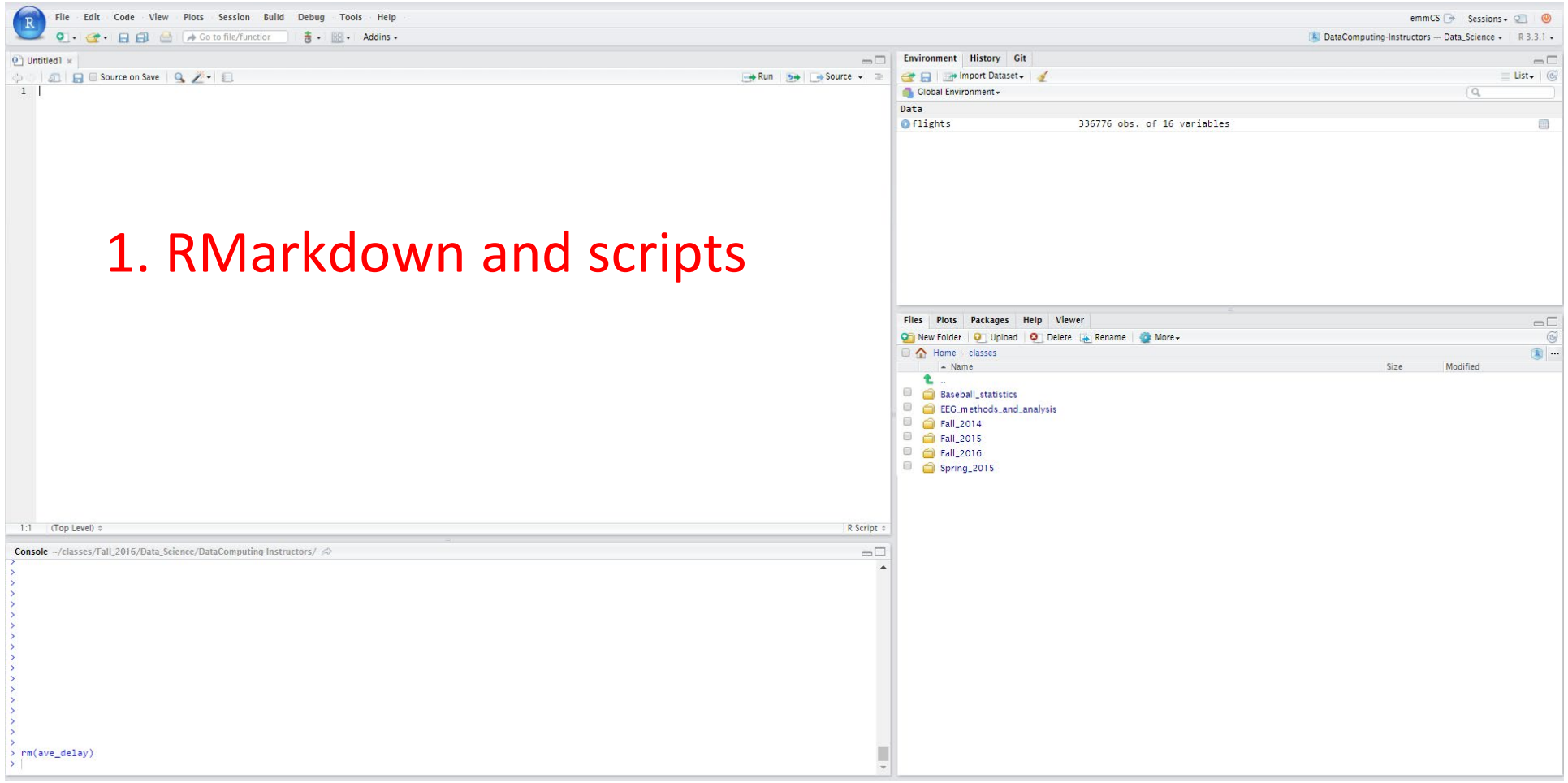
RStudio layout



R as a calculator

```
> 2 + 2
> 7 * 5
```

RStudio layout



<u>Create a new script</u>

File -> New File -> R Script

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

R Basics

Arithmetic:

```
> 2 + 2
> 7 * 5
```

Assignment of values to ***objects***:

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

Number journey…

Number journey

```
> a <- 7
> b <- 52
> d <- a * b
> 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

```
> z > 3
```

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 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?