
Data
Exploration
and
Analysis

Overview

Introductions

Overview and logistics of the course

Review of a few central concepts from Intro Stats

Introduction to R

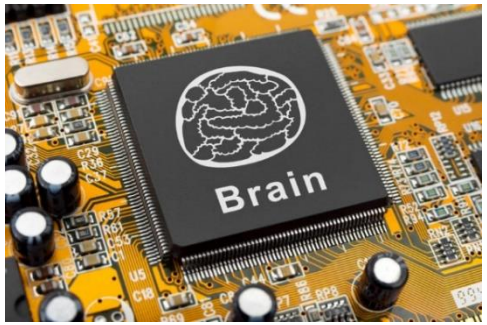
- R as a calculator
- Objects and vectors
- Installing the class SDS230 package and LaTeX (if there is time)


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)

- Amanda Weiss: amanda.weiss@yale.edu

Undergraduate Learning Assistants (ULA)

- Nathan Kim: nathan.kim@yale.edu
- Stephan Billingslea: stephan.billingslea@yale.edu
- Tai Michaels: tai.michaels@yale.edu
- One or two more ULAs who will be joining the class as well soon

Course manager

- William Pang: William.pang@yale.edu

Introductions

Let’s do some quick introductions

Create groups of 3-5 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


What is this class about?

Course objectives

Extend and solidify concepts and method learned in intro stats

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

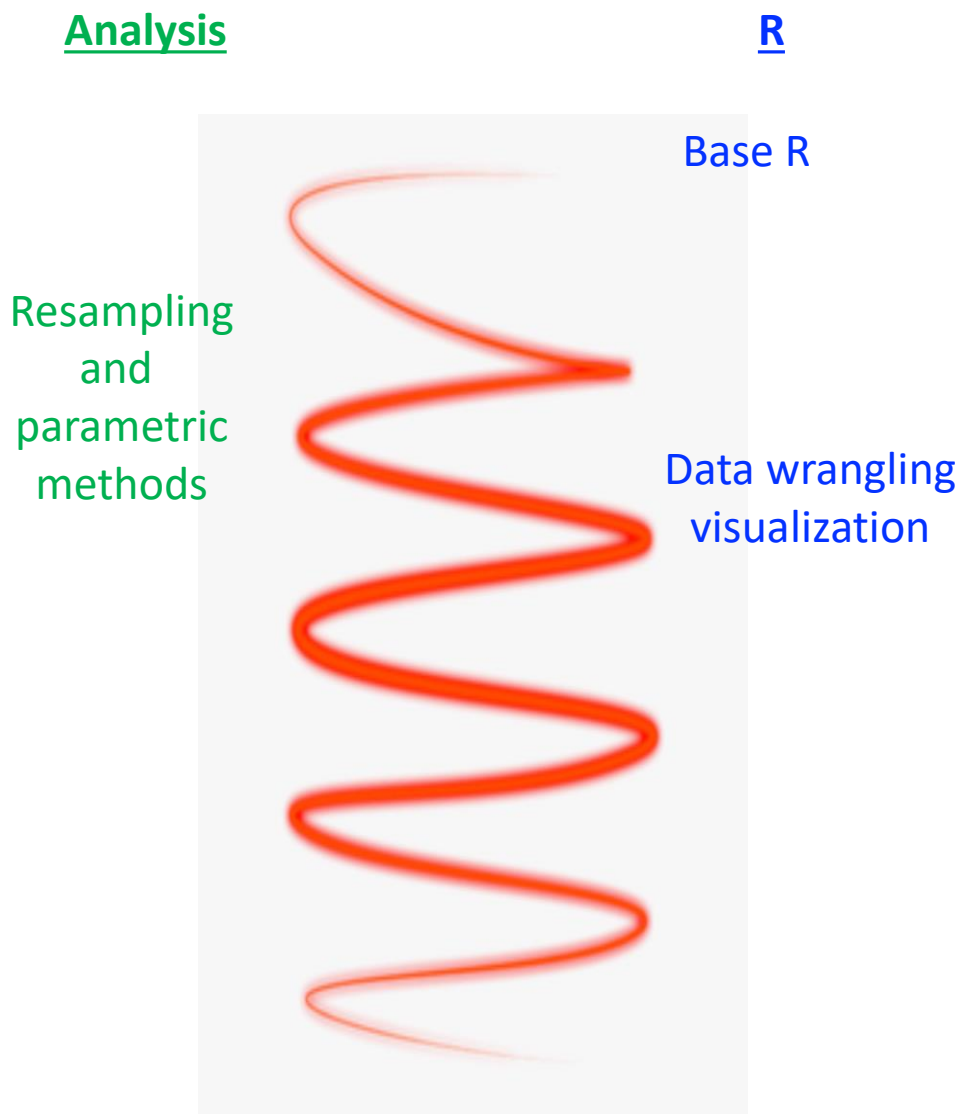
Learn how to use the R programming language to analyze, visualize, and wrangle data

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!

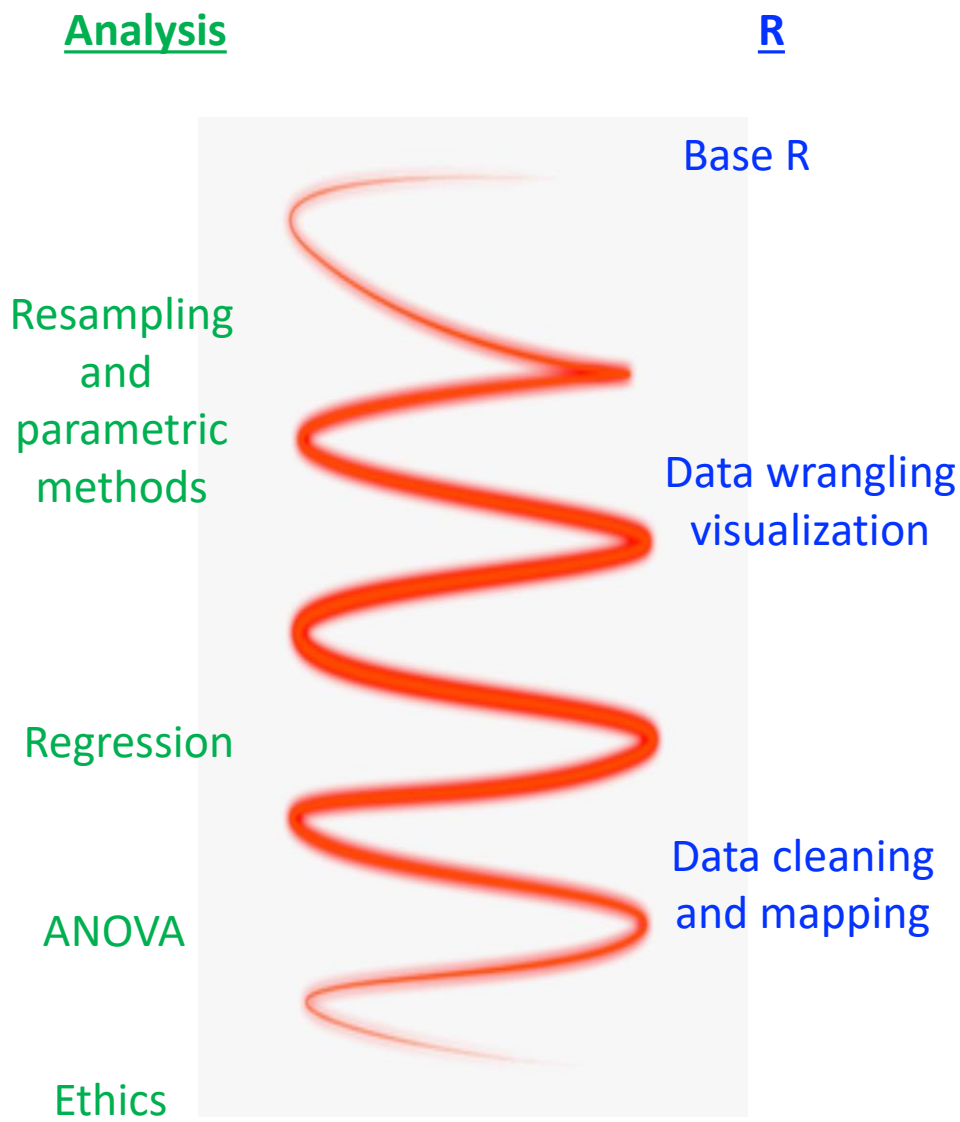
Plan for the semester

1	Sep 1	Course overview, introduction to R, descriptive statistics
2	Sep 6-8	Review of central statistical concepts and exploratory analysis using R
3	Sep 13-15	Confidence Intervals and the bootstrap
4	Sep 20-22	Review of hypothesis tests and permutation tests in R
5	Sep 27-29	Parametric tests and theories of hypothesis testing
7	Oct 11-13	Review and midterm exam
8	Oct 18-23	October break



Plan for the semester

9	Oct 25-27	Simple linear regression
10	Nov 1-3	Multiple regression
11	Nov 8-10	Model selection and logistic regression
12	Nov 15-17	Data cleaning (I'm away at a conference), Joining and mapping
13	Nov 18-27	November break
14	Nov 30 Dec 2	Analysis of Variance (ANOVA)
15	Dec 6-8	Ethics, wrap up and review
16	Dec 19th	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

ANOVA: one-way/factorial, interactions, (mixed effects?)

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

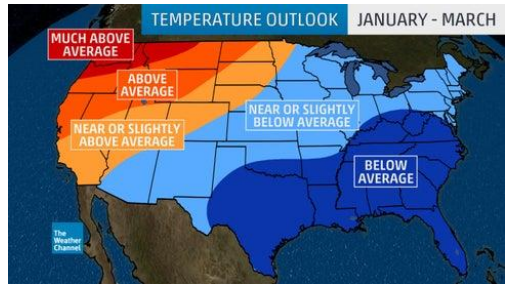
Examples of questions we might look at…

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!

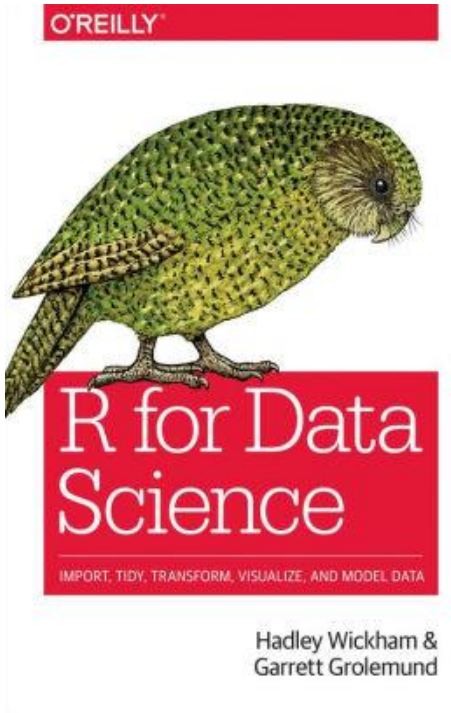
Class structure

Class time 9-10:15am Tuesdays and Thursdays

- New content introduced, questions abswers
- Might have some pre-recorded video content to be watched prior to class

Canvas website:
https://yale.instructure.com/courses/79947

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



Office hours

My planned office hours (subject to change)

- Tuesday and Thursday at 11am
- 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

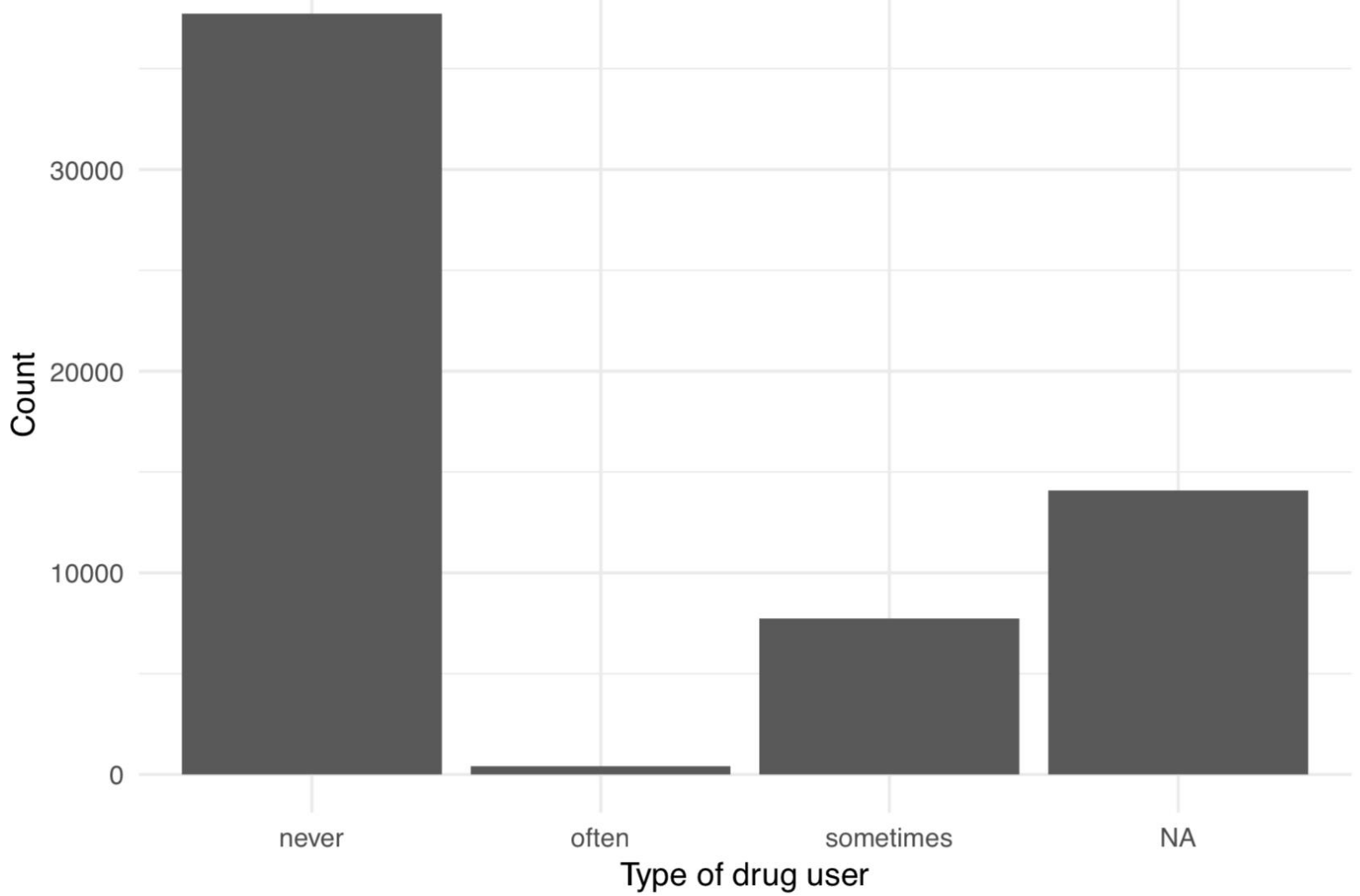
Assignments and grades

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

Homework policies

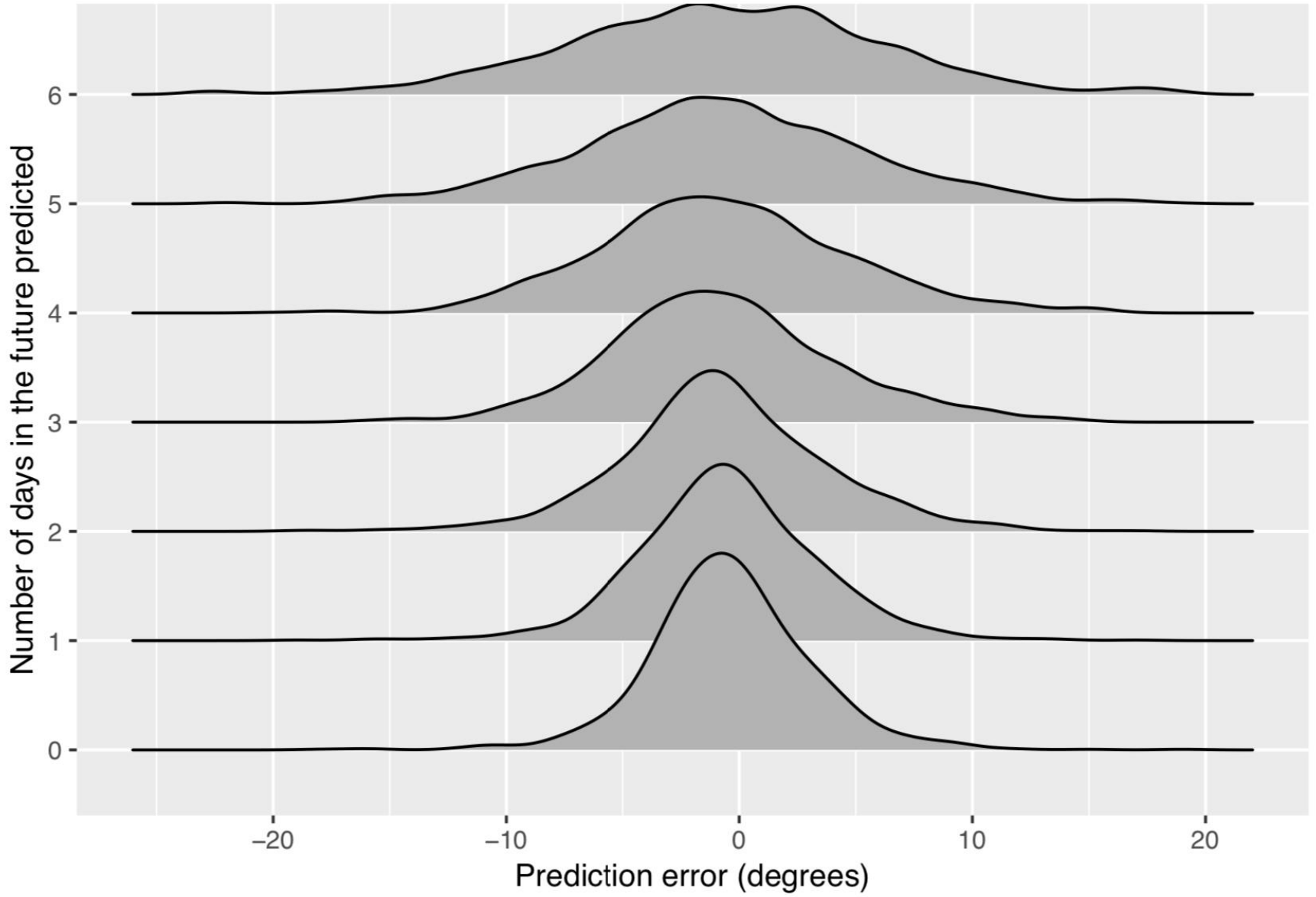
- You may discuss questions with other but the work you turn in must be your own
- Homework assigned on Tuesdays and are due at 11pm on Sundays
  - (with a 59 minute grace period)
- Late worksheets (90%) credit if turned in by 11:59pm on Monday
  - For any other extension a Dean's Extension is needed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

Example 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("")
```

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).

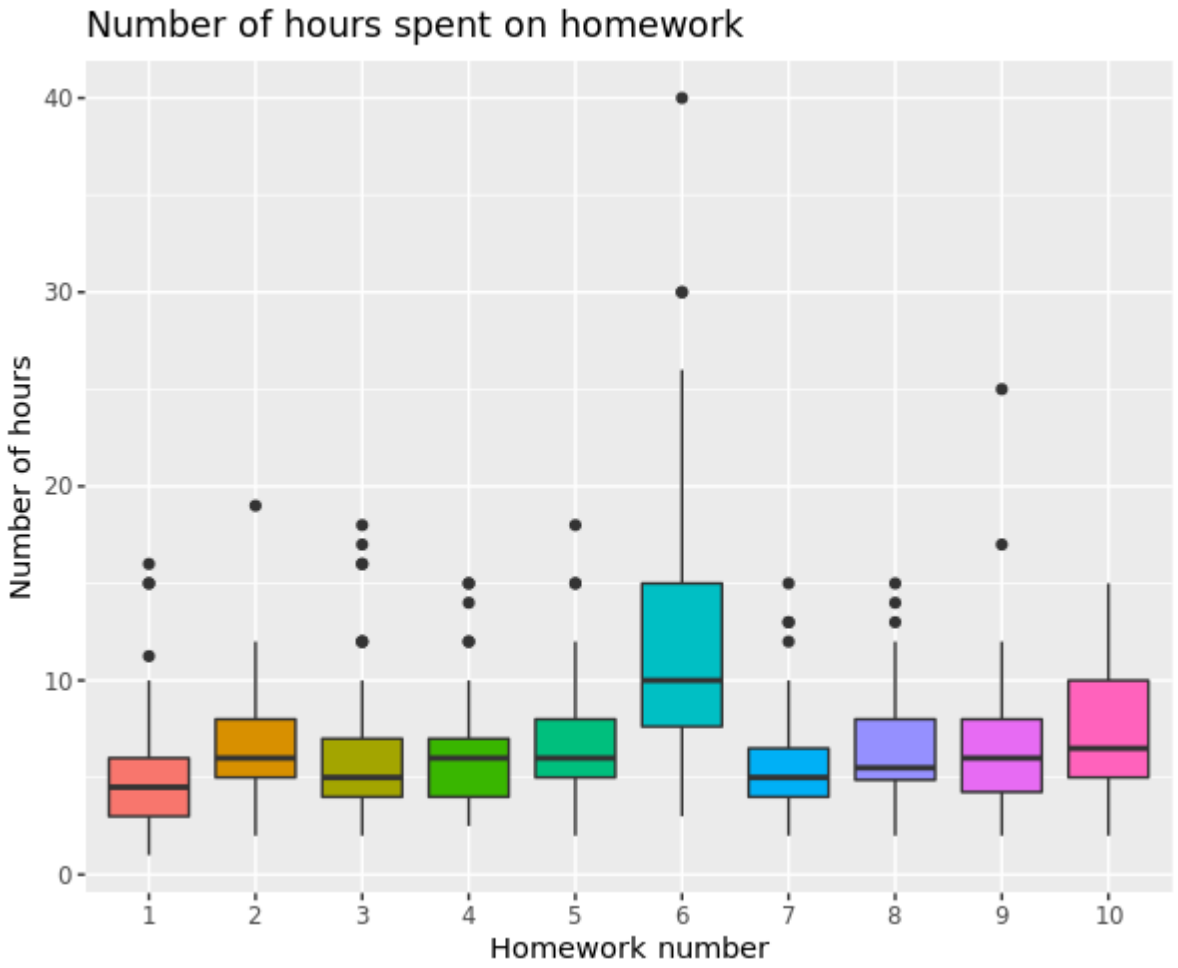
Assignments and grades

2. Final project (8%)
   - Find a data set and analyze it on your own (5-7 page report)

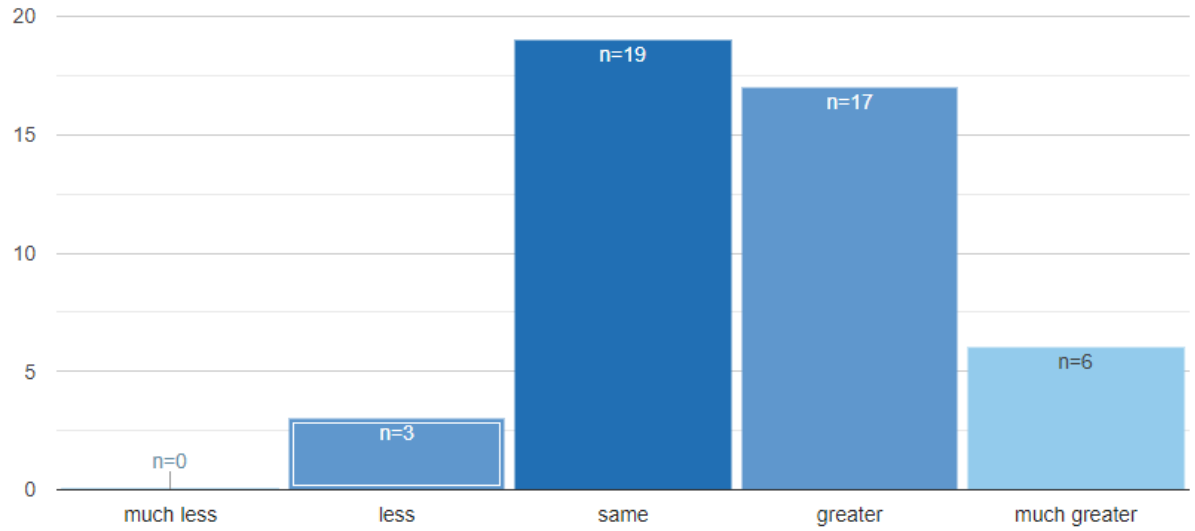
3. Exams (36% total)
   - Midterm (15%) Oct 14th during class
   - Final (21%) Dec 21st at 7pm

4. Participation (2%)
   - Active asking and answering questions on Ed Discussions
     - Full credit will be given for 8 or more questions or answers

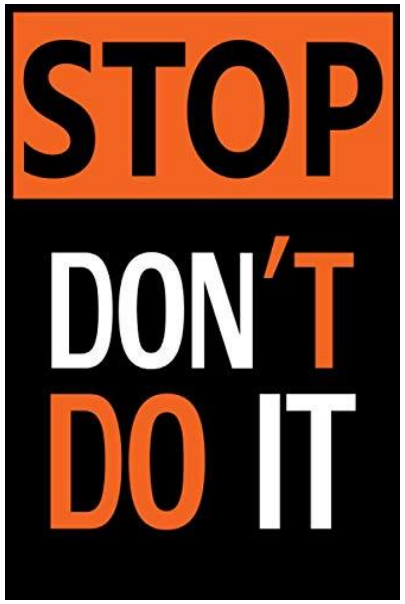
How much work is this class?



Question Average	Dept	Div	School
3.6 ±0.8	3.1 ±0.9	3.3 ±1.0	3.1 ±1.0



Academic honesty



Plagiarism/cheating

- Yale's Academic Integrity Statement

You are allowed to talk with others about the homework, but the work you turn in must be your own

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

Class background survey

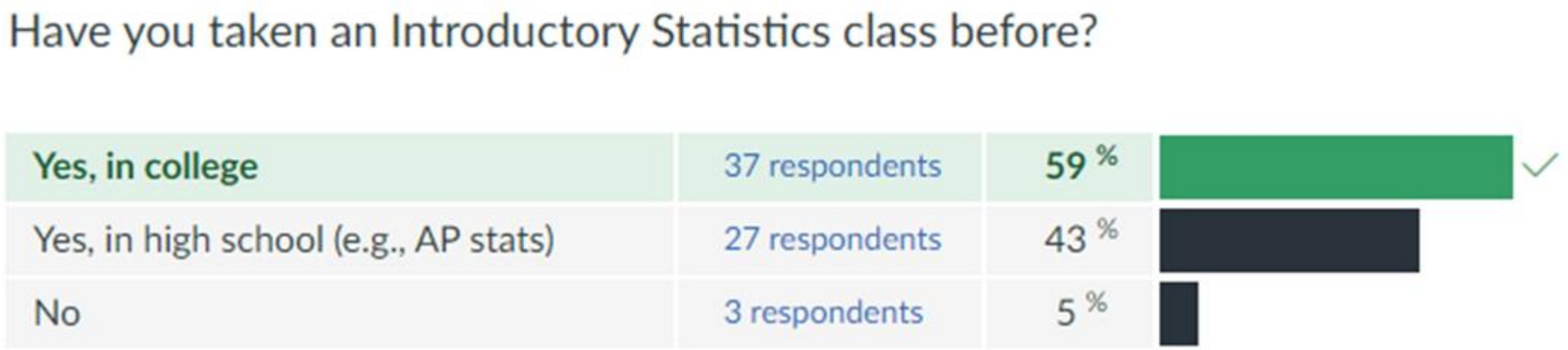
In order for me to get to know you and to better adjust the class to your interests, please fill out the class background survey on canvas

- Under the Quizzes link on the left on Canvas

Preliminary class survey results

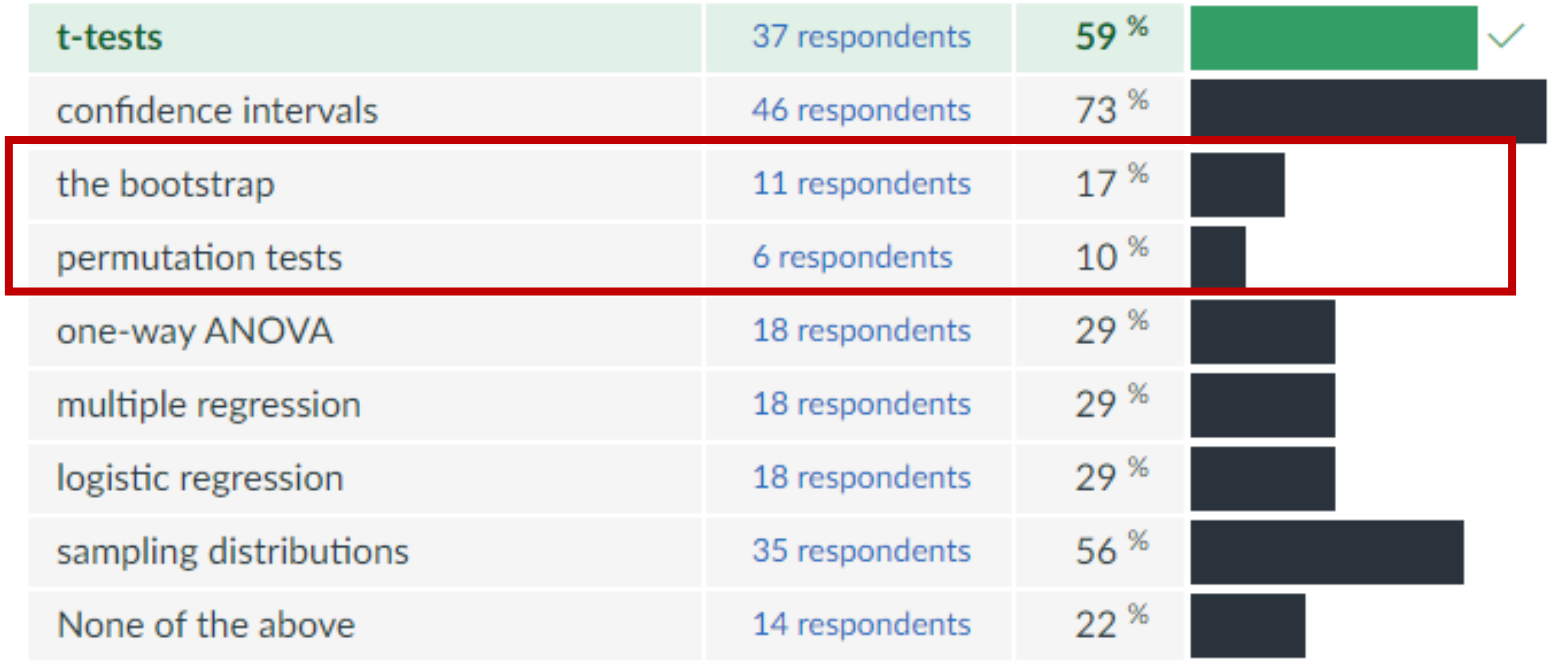
As of 6:30pm yesterday, 63 people had filled out the class survey

- ~50% of the class is undergraduates, ~50% graduate students



Class survey results

Which Statistics methods/concepts are you comfortable with?



t-tests	37 respondents	59 %
confidence intervals	46 respondents	73 %
the bootstrap	11 respondents	17 %
permutation tests	6 respondents	10 %
one-way ANOVA	18 respondents	29 %
multiple regression	18 respondents	29 %
logistic regression	18 respondents	29 %
sampling distributions	35 respondents	56 %
None of the above	14 respondents	22 %

Class survey results

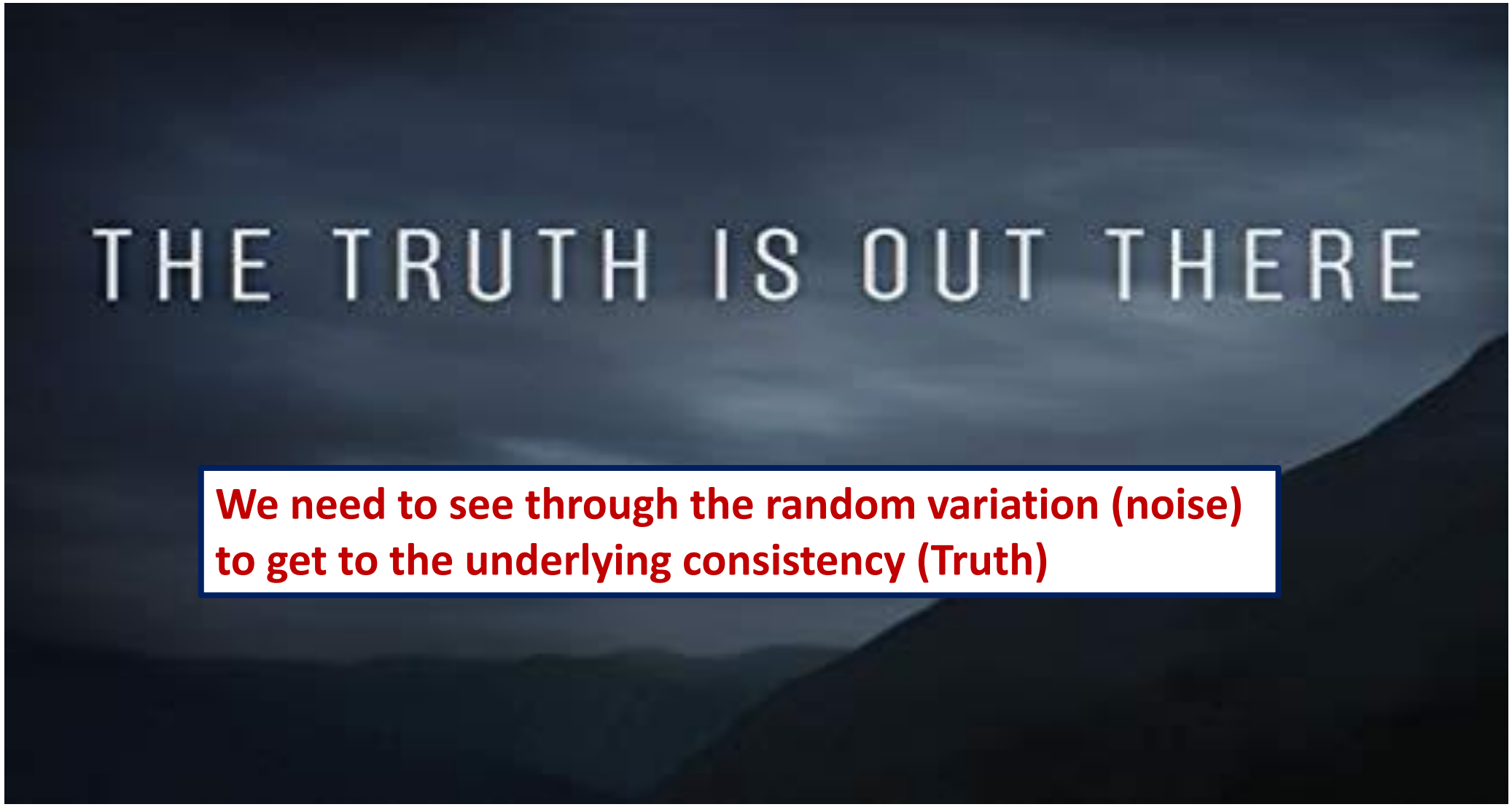
How much experience do you have with computer programming?

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

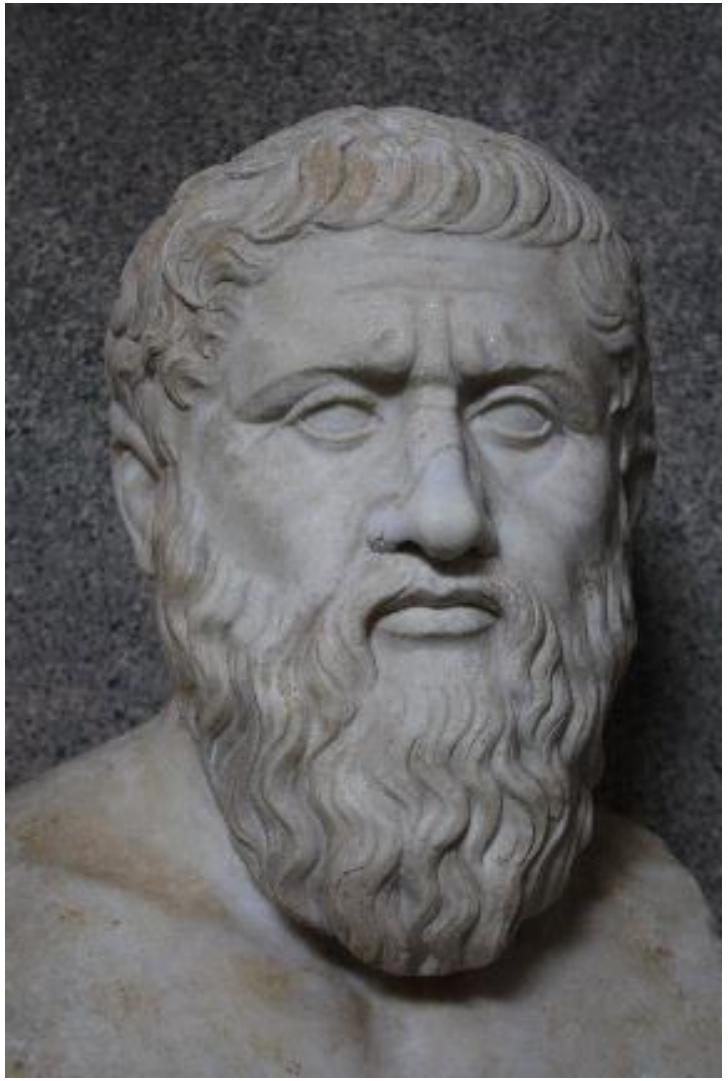

ANY QUESTIONS?

Quick Review of central concepts in Intro Statistics

Quick Review of central concepts in Intro Statistics



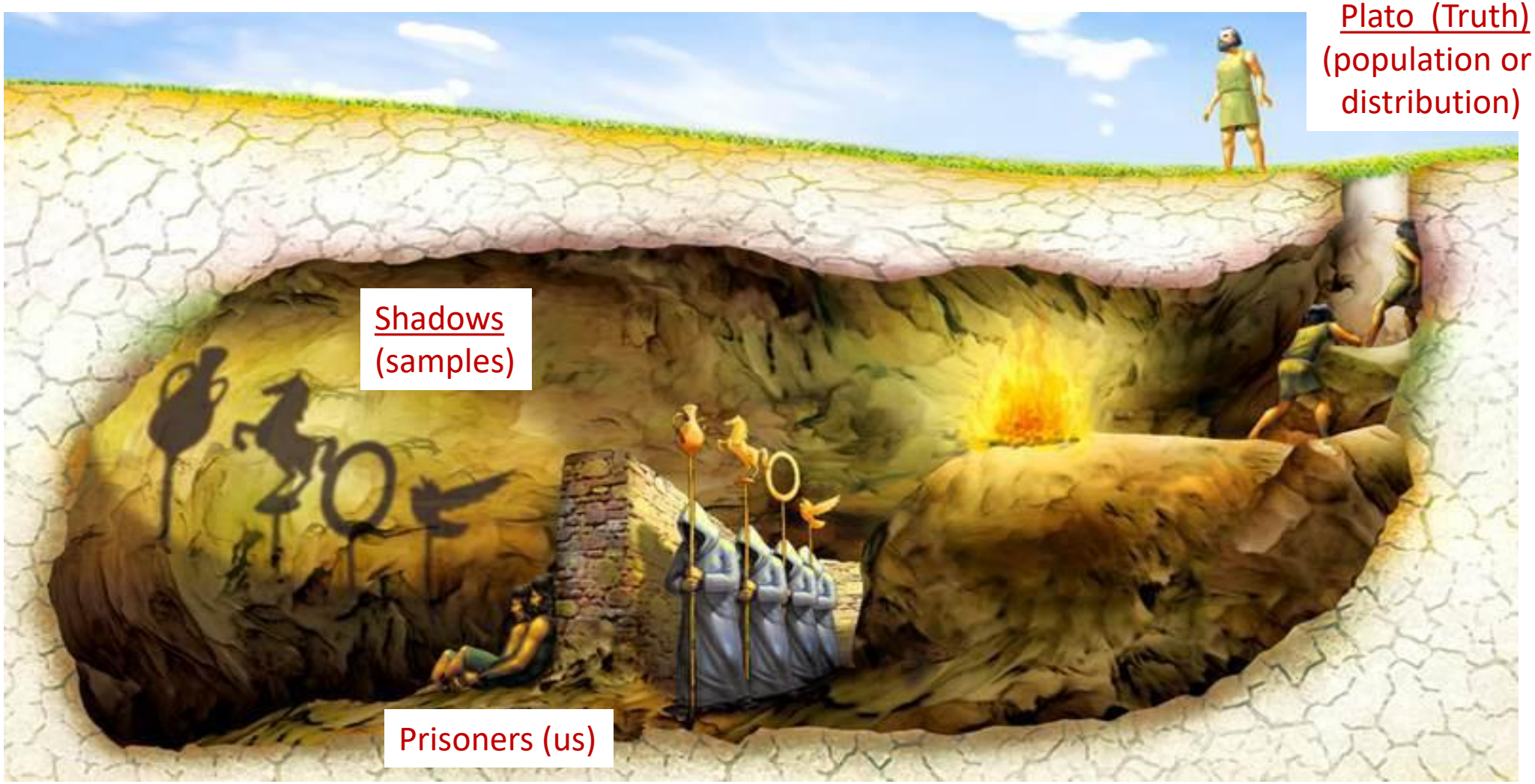
The Truth®!



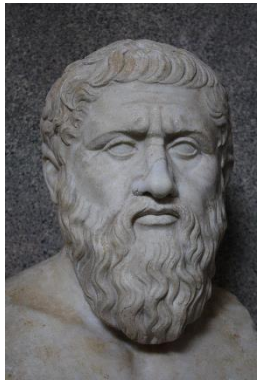
If we could see all the (infinite) data, we would know the Truth®!

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

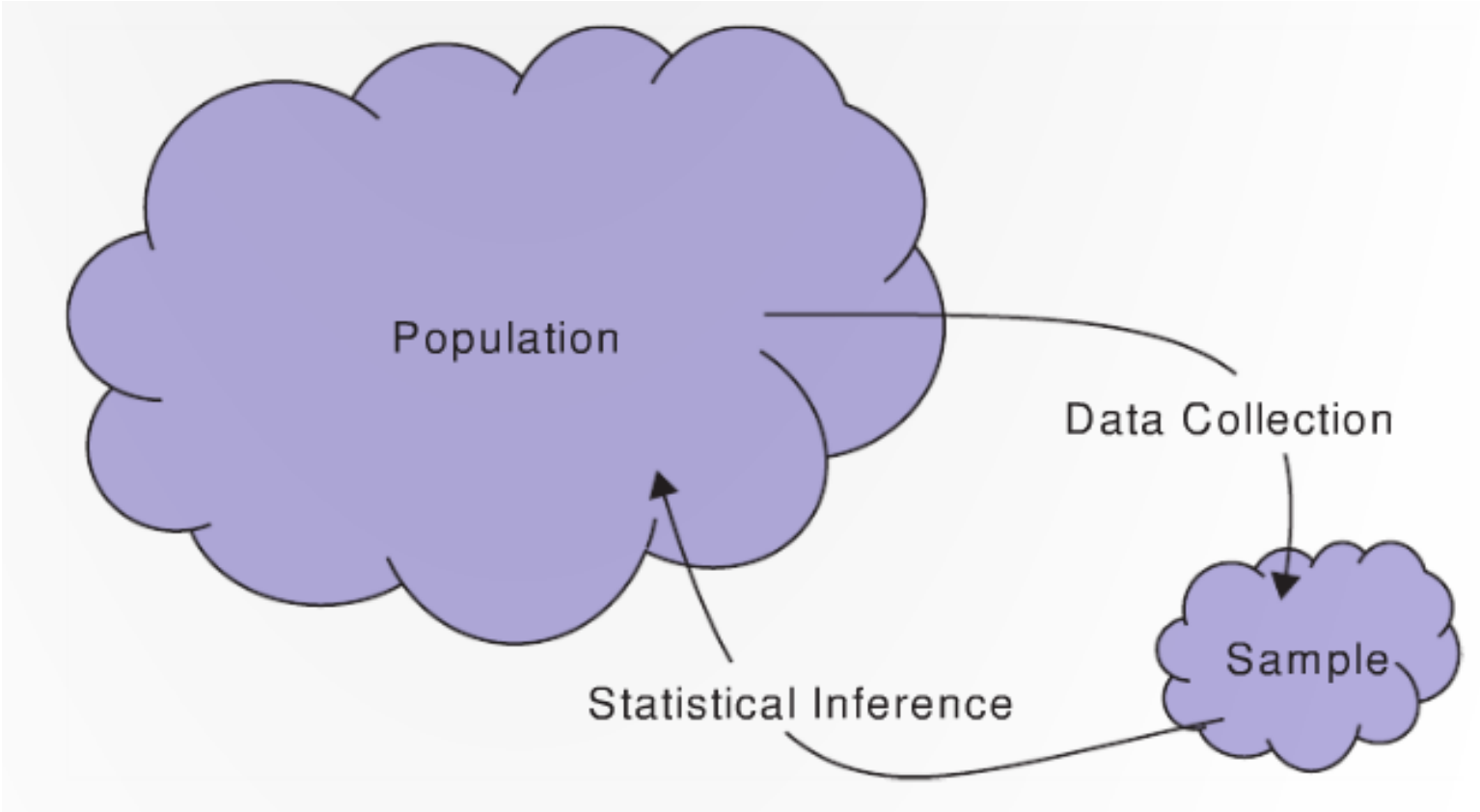
Plato's cave

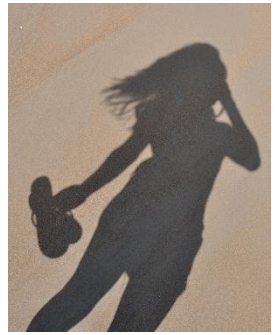


From The Republic (~ 380 BCE)

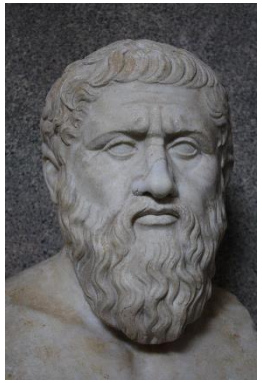


Population: all individuals/objects of interest



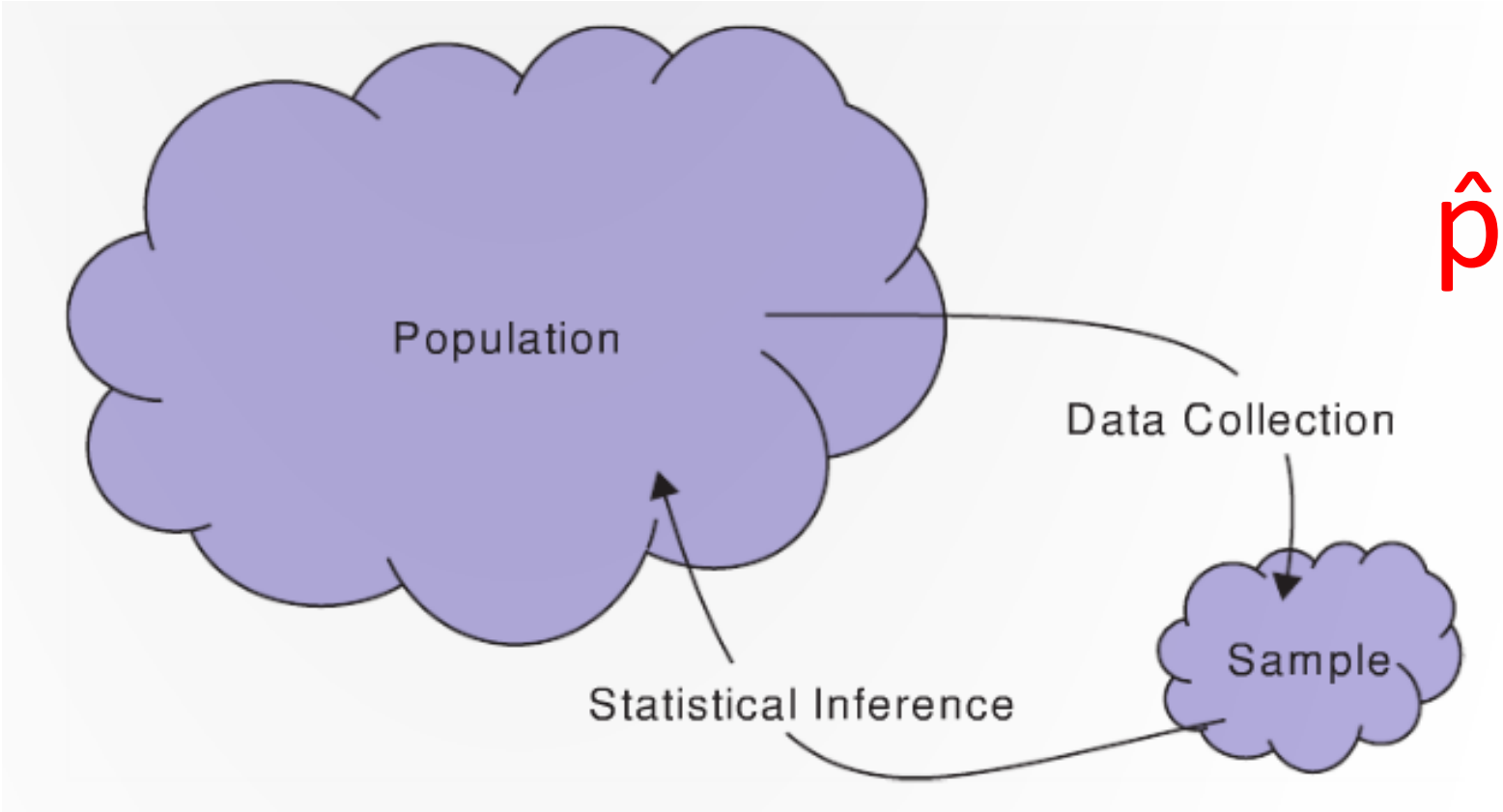


Sample: A subset of the population

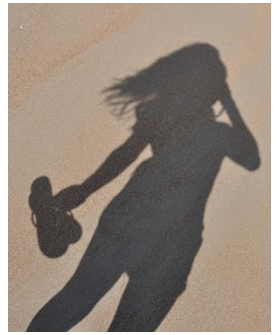


$\pi, \mu, \sigma, \rho, \beta$

Parameter: a number characterizing a property of a population

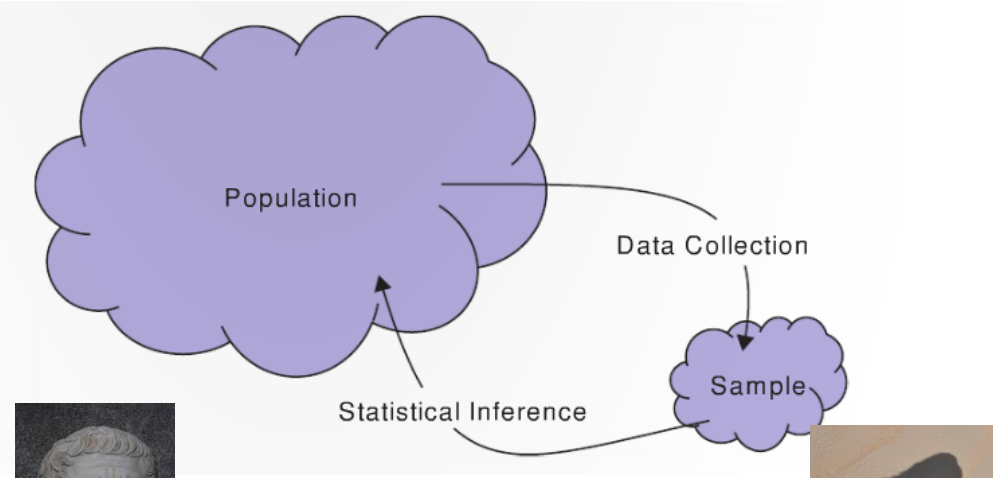


$\hat{p}, \bar{x}, s, r, b$

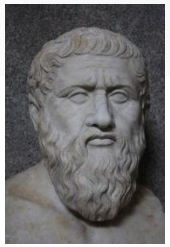


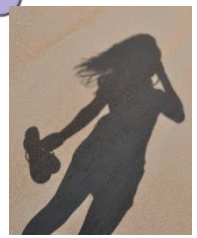
Statistic: A number computed from a sample

Parameters and statistics commonly used symbols



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

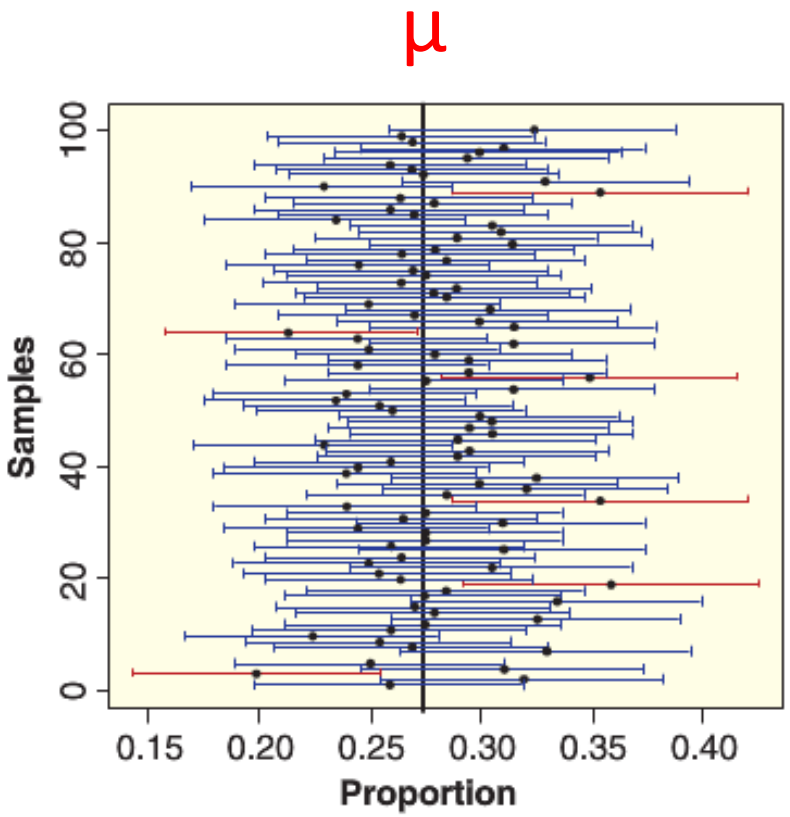




	Population parameter (Plato)	**Sample statistic (shadow)**
Mean		
Standard deviation		
Proportion		
Correlation		
Regression slope		

Inference on parameters

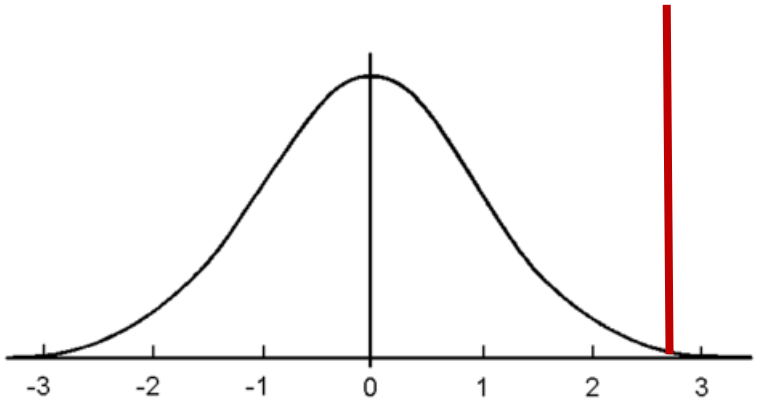
Confidence intervals



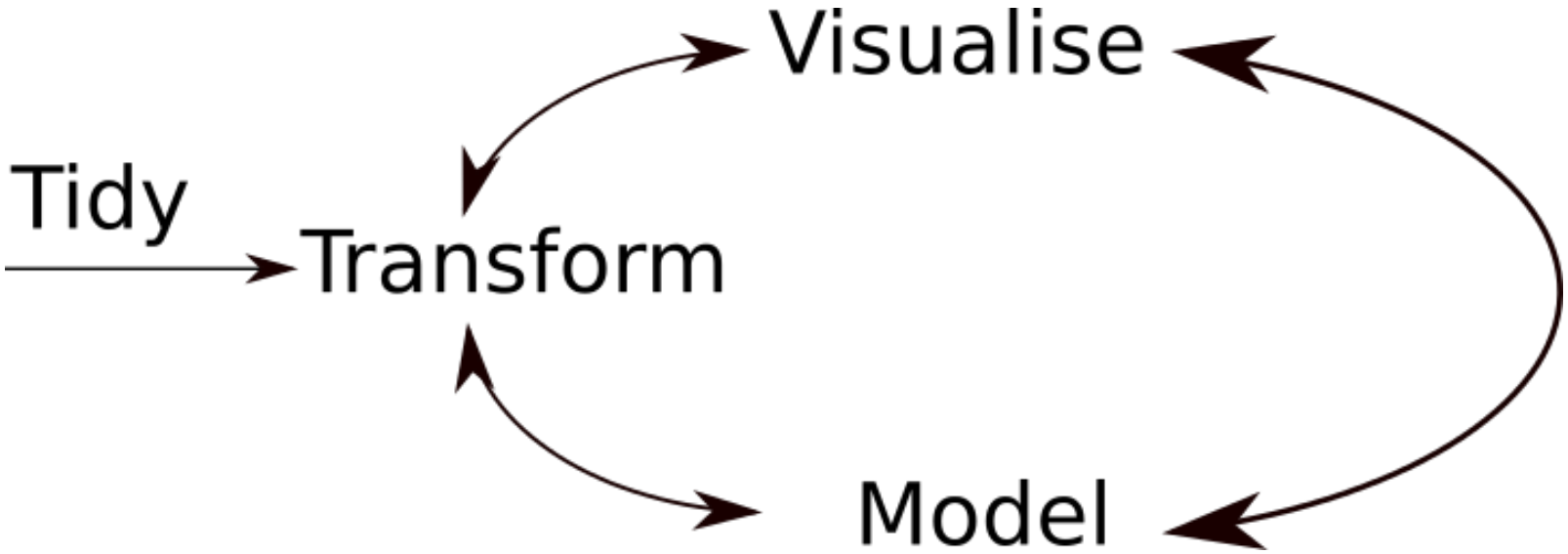
Hypothesis tests

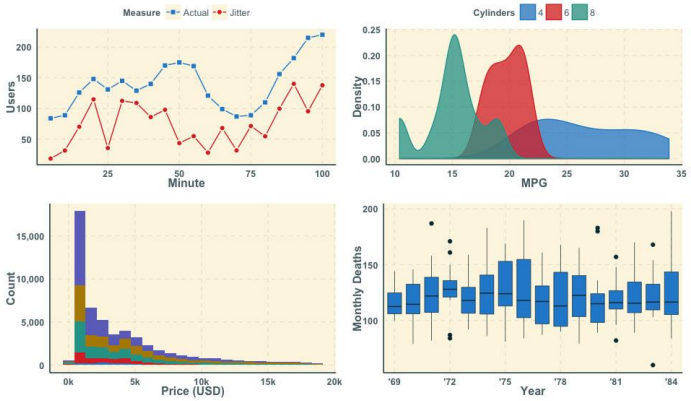
H_0: $\mu = 0$

H_A: $\mu > 0$

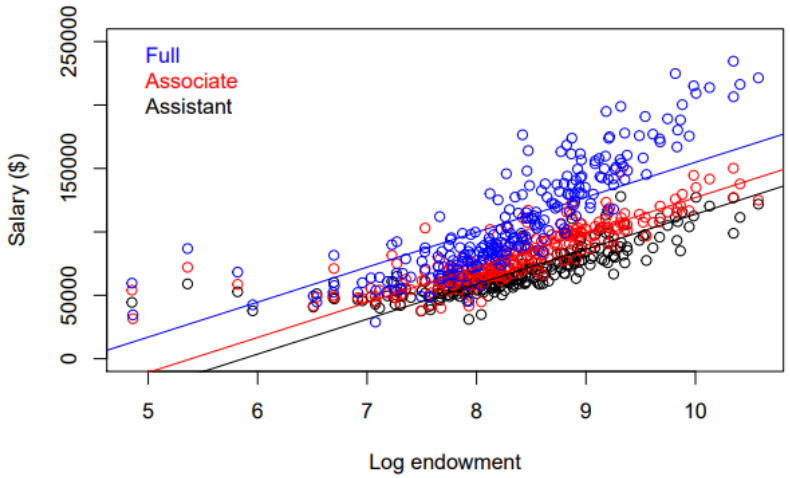


Sometimes the Truth is more complicated…





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



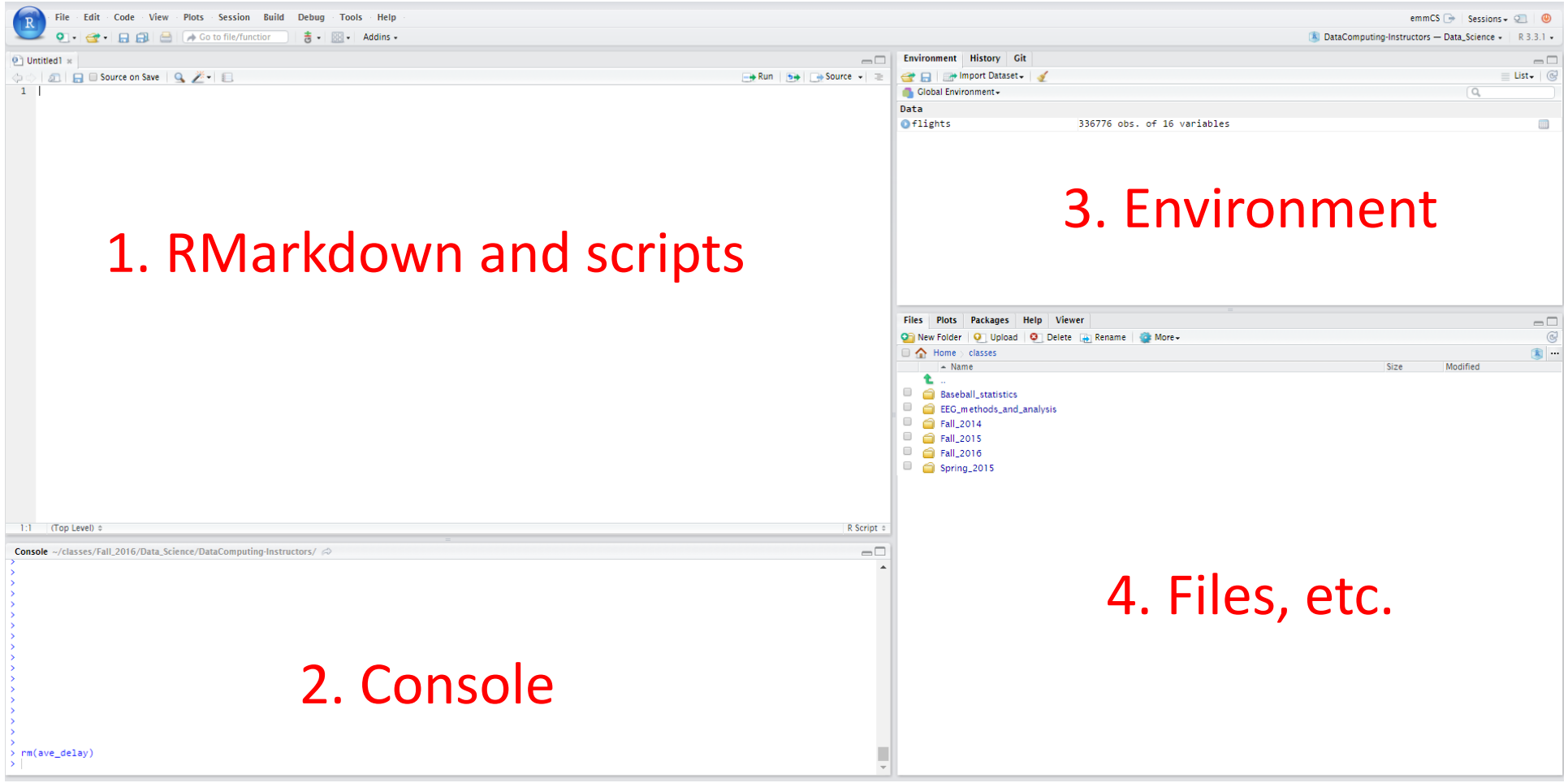

Questions?

R and R Studio

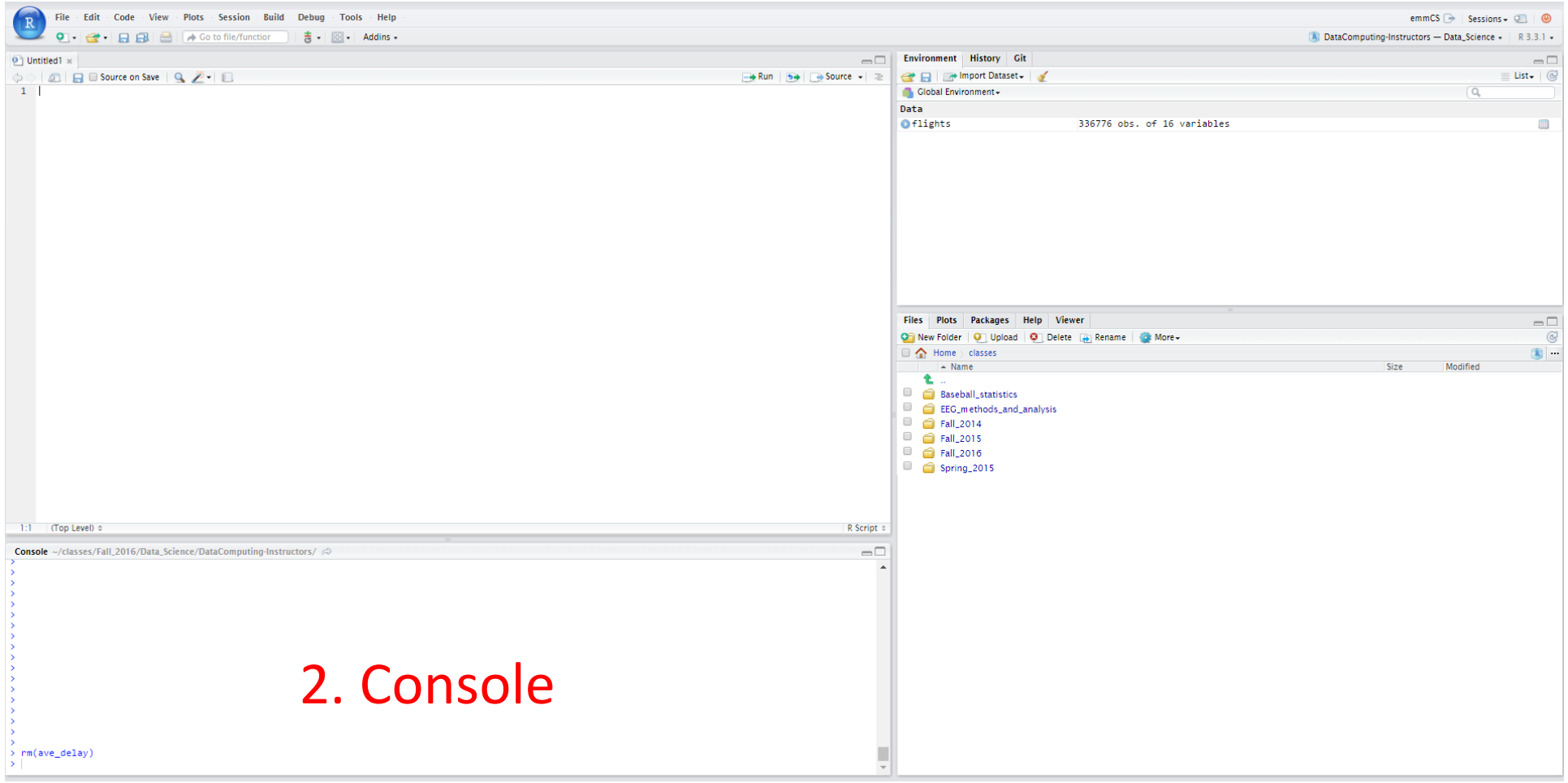
R: Engine

RStudio: Dashboard

RStudio layout



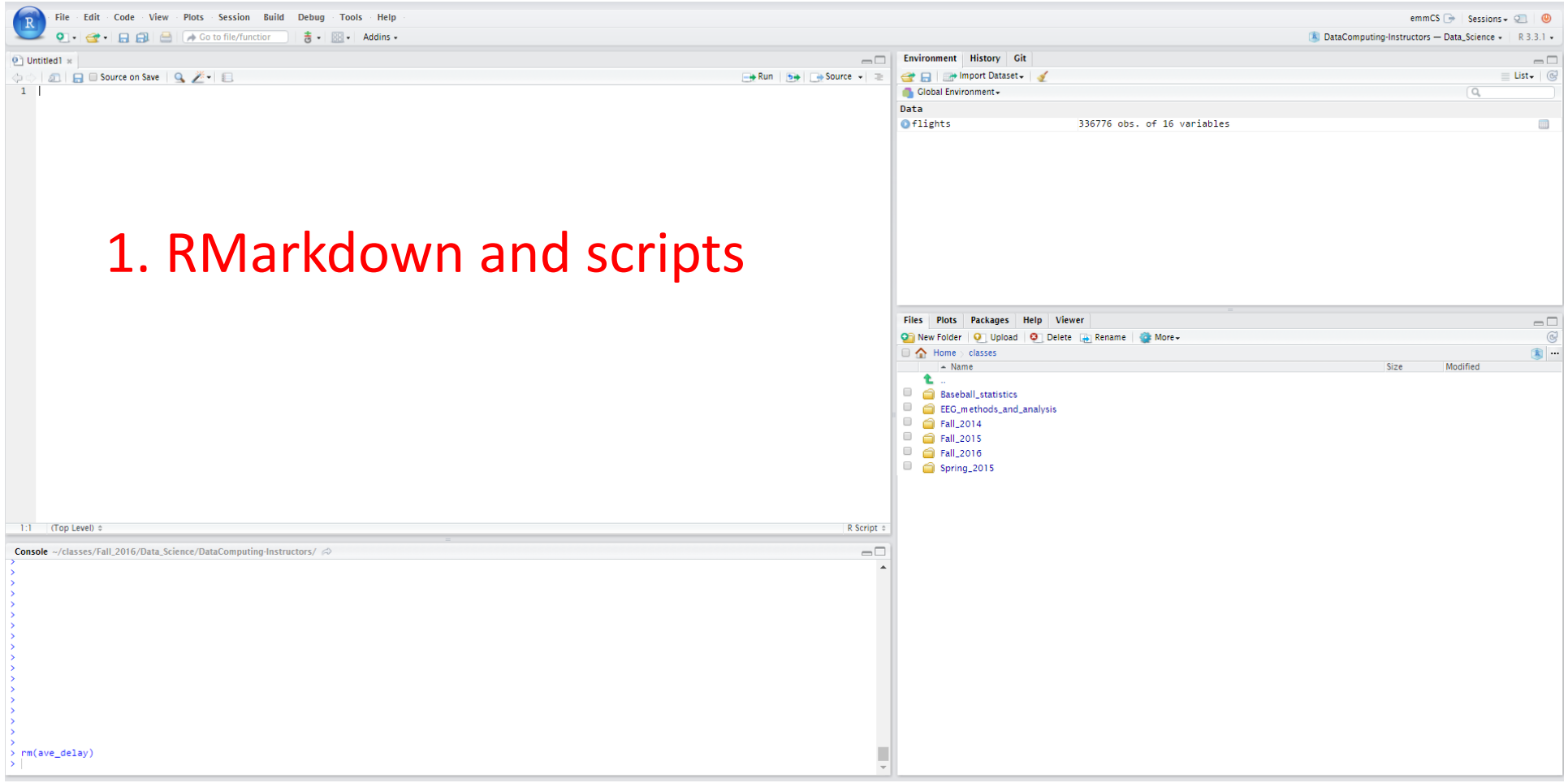
RStudio layout



R as a calculator

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

RStudio layout



1. RMarkdown and scripts

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 + 2
> 7 * 5
```

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

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

Number journey…

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?