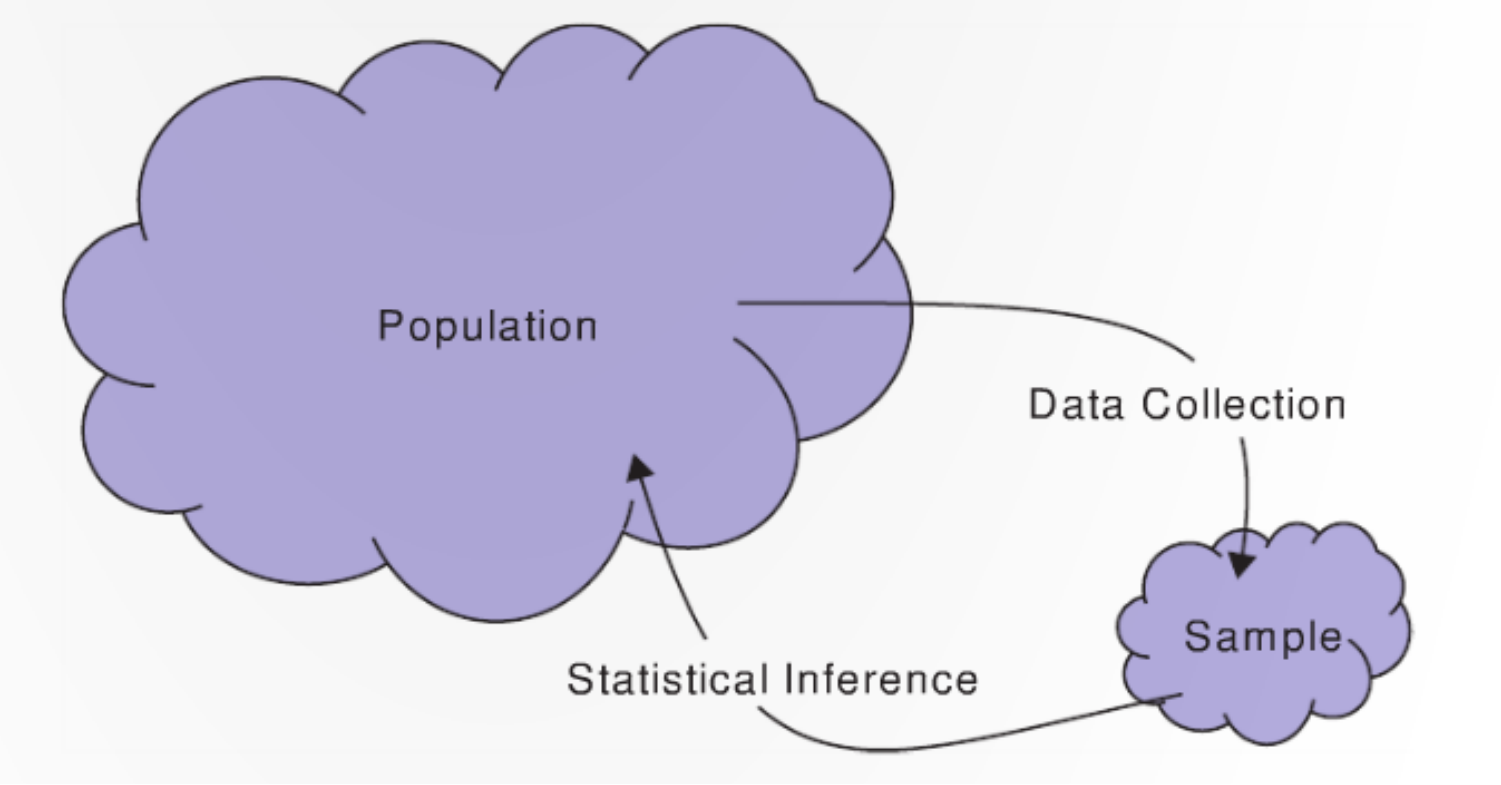
Class 1: logistics and central concepts in Statistics



Course overview and logistics

Office hours and contact information

Email: ethan.meyers@yale.edu

Office hours: 10:45-11:45am Tuesdays and Thursdays

Office: Kline Tower, room 1253

- https://yale.zoom.us/j/92321390756

Teaching Assistants

Teaching Fellow

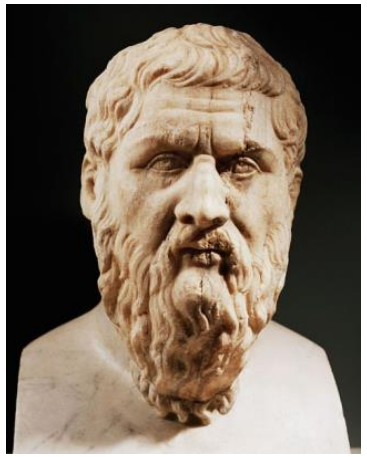
- Tom Shin: tom.shin@yale.edu

Undergraduate Learning Assistants

- Lydia Monk: lydia.monk@yale.edu
- Cindy Nguyen: cindy.nguyen@yale.edu
- Elizabeth Greenberg: elizabeth.greenberg@yale.edu
- Alexandre Vantassel: alexandre.vantassel@yale.edu
- Eve Cohen: eve.cohen@yale.edu
- Clara Kleindorfer: clara.kleindorfer@yale.edu

Learning goals

1. Understand the key concepts in Statistics



2. To learn how to analyze real data
   - We will use the R programming language
     - Do not fear, this will make our life easier!

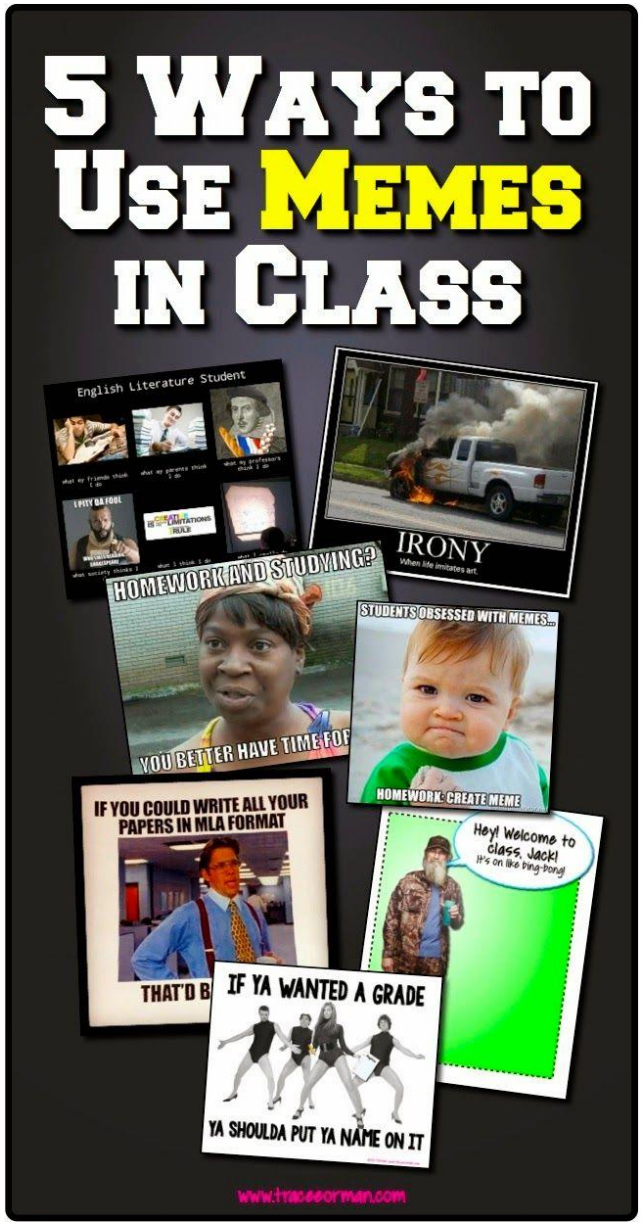
Course format

Pre-recorded videos and quizzes/surveys will be available prior to the regularly scheduled class time

- Ideally watch them in advance of scheduled class

The regularly scheduled class time (9-10:15am Tues/Thurs) will be an opportunity to:

- 9-9:30 Review conceptual material and ask questions
- 9:30-10:15: Practice additional exercises and analyses



Plan for the semester

<u>Exploring data/descriptive statistics (weeks 1-4)</u>

Sampling, categorical and quantitative data

Measures of central tendency and spread

- Mean, median, standard deviation

Relationships between variables

- Correlation and regression

Plan for the semester

Inferential Statistics

Sampling distributions

Confidence intervals

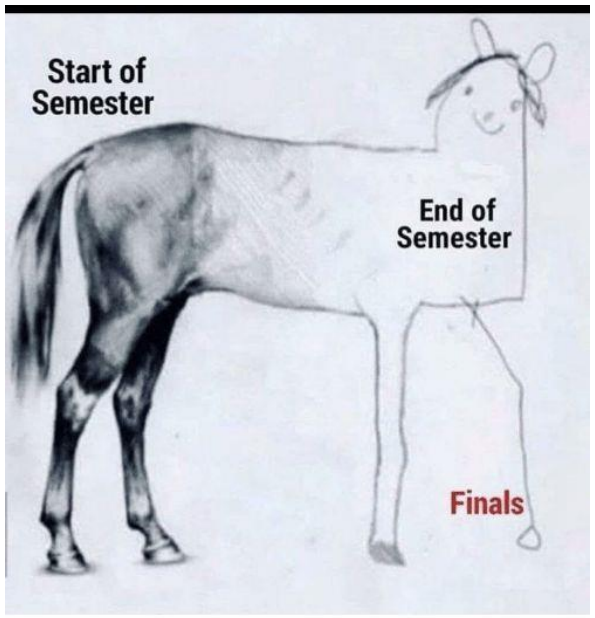
- The bootstrap

Hypothesis tests using randomization methods

- Permutation tests

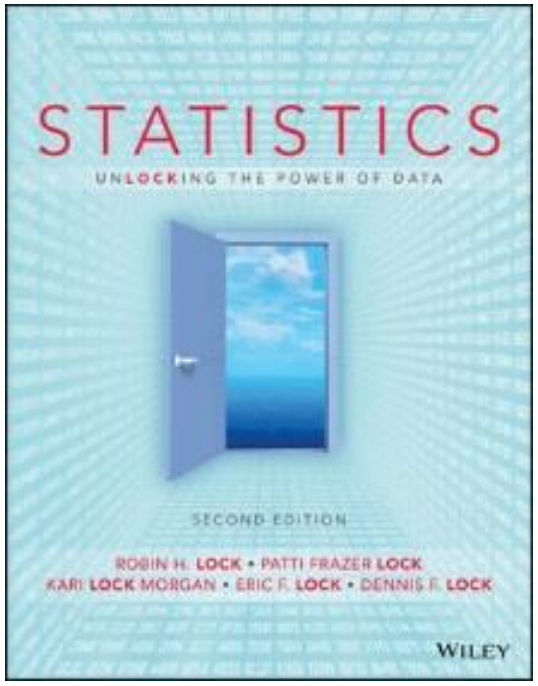
Hypothesis tests using parametric methods

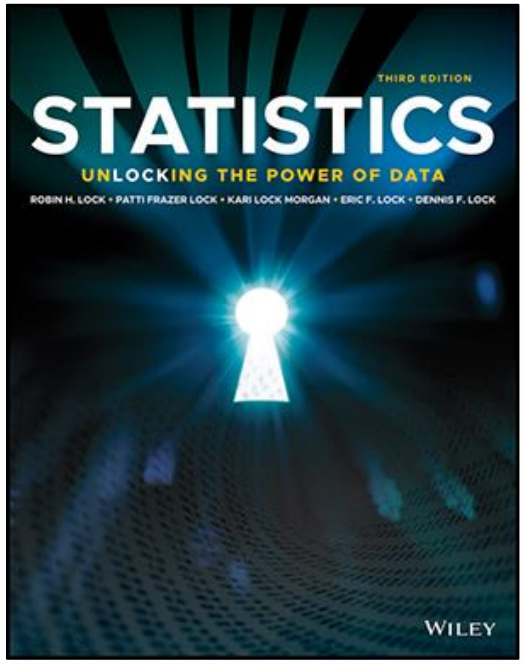
- T-tests, ANOVA, etc.

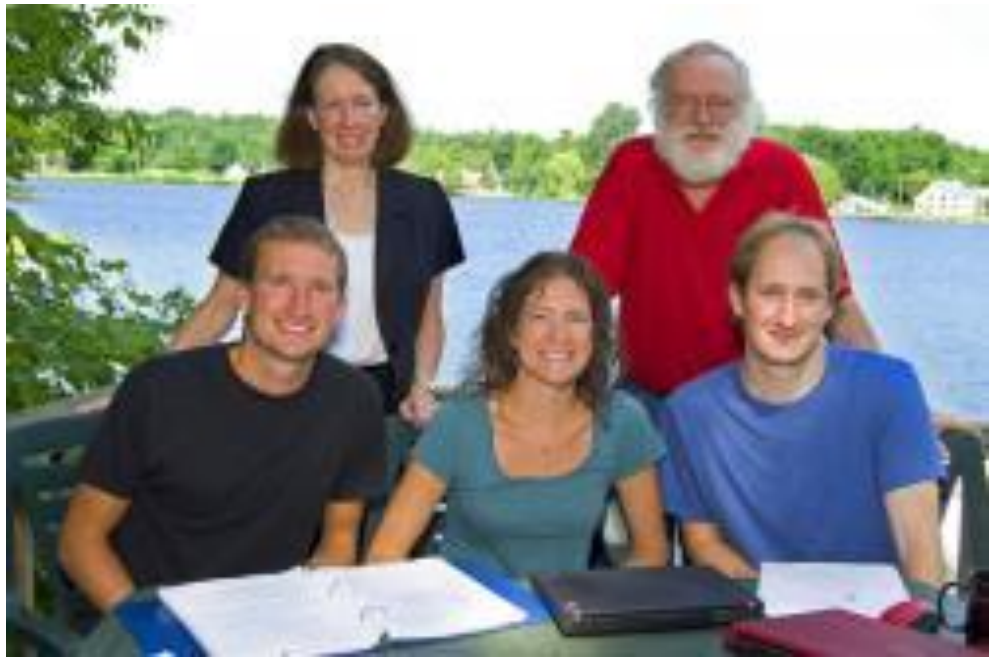


me irl

Textbook: Lock5

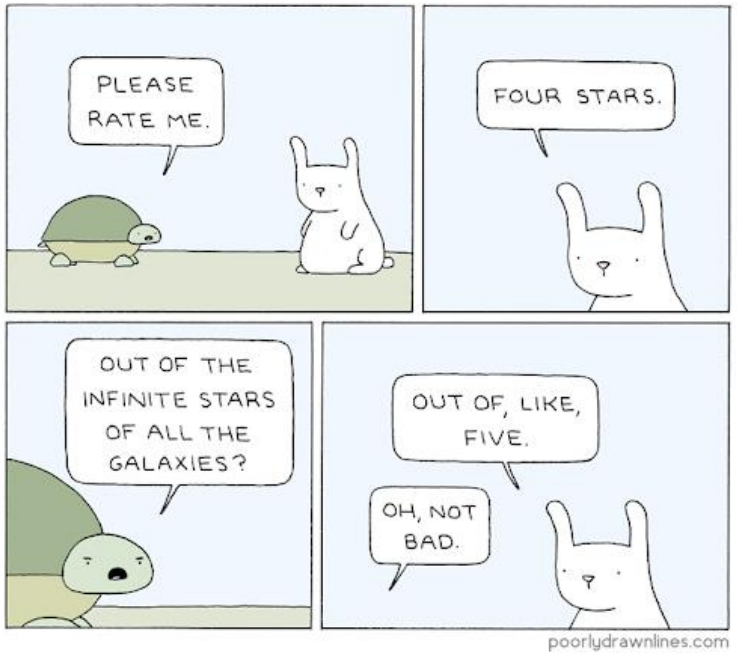






Additional reading and other resources will be posted to Canvas: https://yale.instructure.com/courses/95067

Assignments and grades



1. Homework problem sets (45%)
   - Exploring concepts and analyzing data using R
   - Weekly: 10 in 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 11pm on Monday
  - For any other extensions a Deans Excuse is needed
- Lowest scoring worksheet will be dropped

Assignments and grades

2. Final project (10%)
   - Similar in length to a homework assignment, but you will analyze data of your own choosing based on your interests using methods discussed in the class

3. Exams (43% total)
   - Midterm: March 7th during the regular class time (15%)
   - Final: May 9th at 9am (28%)

4. Participation (2%)
   - Active asking and answering questions on Ed Discussions

Policies

Accommodation: please let me know if you have accommodations for homework and/or exams

Academic dishonesty: Don’t do it!

- You can work with others on the homework but the work you turn in needs to be your own
  - i.e., you need to understand the concepts and be able to produce the results yourself
- You can’t talk with others on exam, etc.

Examples of questions/analyses we will look at…

Z-scores: What is most impressive about LeBron James?

Sampling: How can insights from the Swedish chef help us avoid bias?

Confidence intervals: How can we pick ourselves up from the bootstrap to estimate a plausible range of values?

Randomization tests: Is it possible to smell whether someone has Parkinson's disease?

A typical homework assignment

Part 2: Practicing R

Please answer the following questions to get practice using a few basic R functions. Make sure you have a clear understanding of how to use this code since future class work will build on this knowledge.

Exercise 2.1: (4 points) Let's get started by using R as a calculator. Use R to calculate the square root of 21.32, and then divide this number by 2.71.

```
# delete the below lines and replace with the correct math (2
# + 3)^2
```

Exercise 2.2: (6 points) Create a vector with the numbers 7, 15, 18, 3, 5, 12, and 20 in it and assign this vector to an object called my_vec. Multiply this vector by 2 and assign it to the object my_vec2. Finally, use the sum() function to sum all the values in the vector my_vec2.

Class survey

In order to get to know you and to adjust the class to everyone's interests, please fill out the class survey on canvas

- Under the Quizzes link on the left

Any questions about the class logistics???

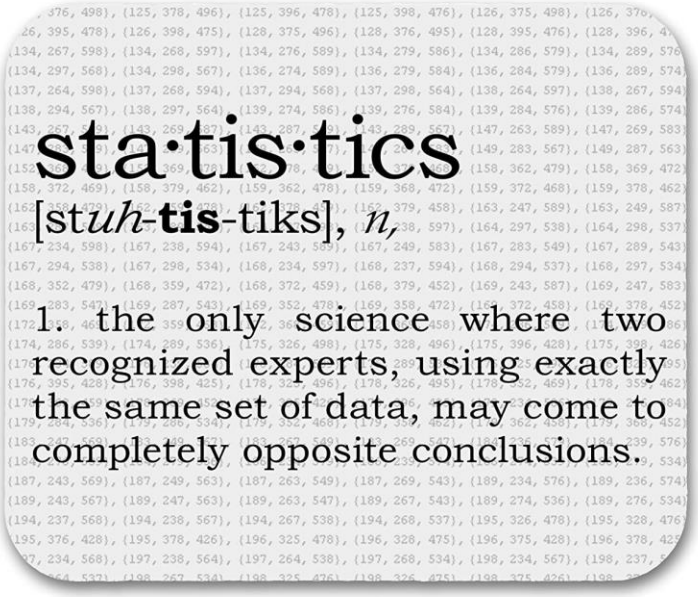
- Ask on Ed Discussions!

What is Statistics?

What is Statistics? (capital S)

"Statistics is a way of reasoning, along with a collection of tools and methods, designed to help us understand the world" (De Veaux et al. 2006, p. 2)

"Statistics is a body of methods for making wise decisions in the face of uncertainty" (Wallis & Roberts 1962, p. 11)

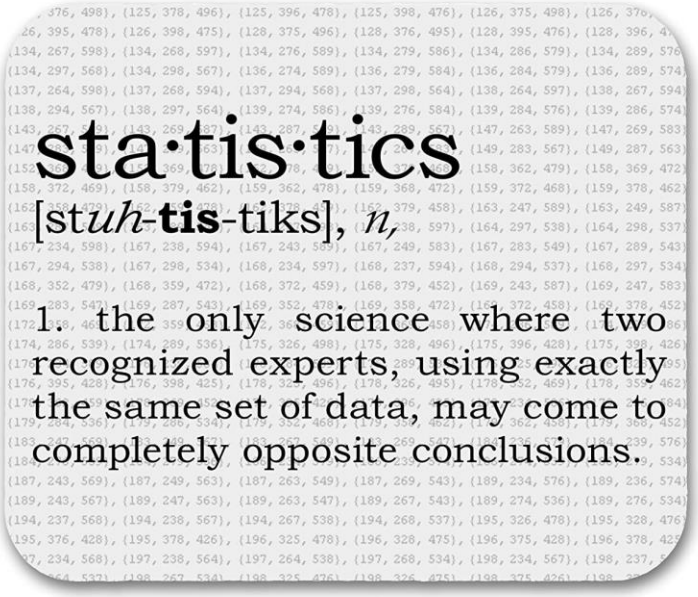


Fienberg, S. (2014). What is Statistics? The *Annual Review of Statistics and Its Application,* 1:1-9

My thoughts

Statistics is a way to use data to answer questions:

- Often we use a small amount of data to answer questions about a larger underlying phenomenon
- We want to know the Truth, and not be fooled by randomness
  - Quantify uncertainty



It's <u>part</u> of an argument

- Don't blindly trust statistical tests, think about the results!
  - Do you really believe them?
- Be your own worst critic and try to prove yourself wrong

A warning about terminology

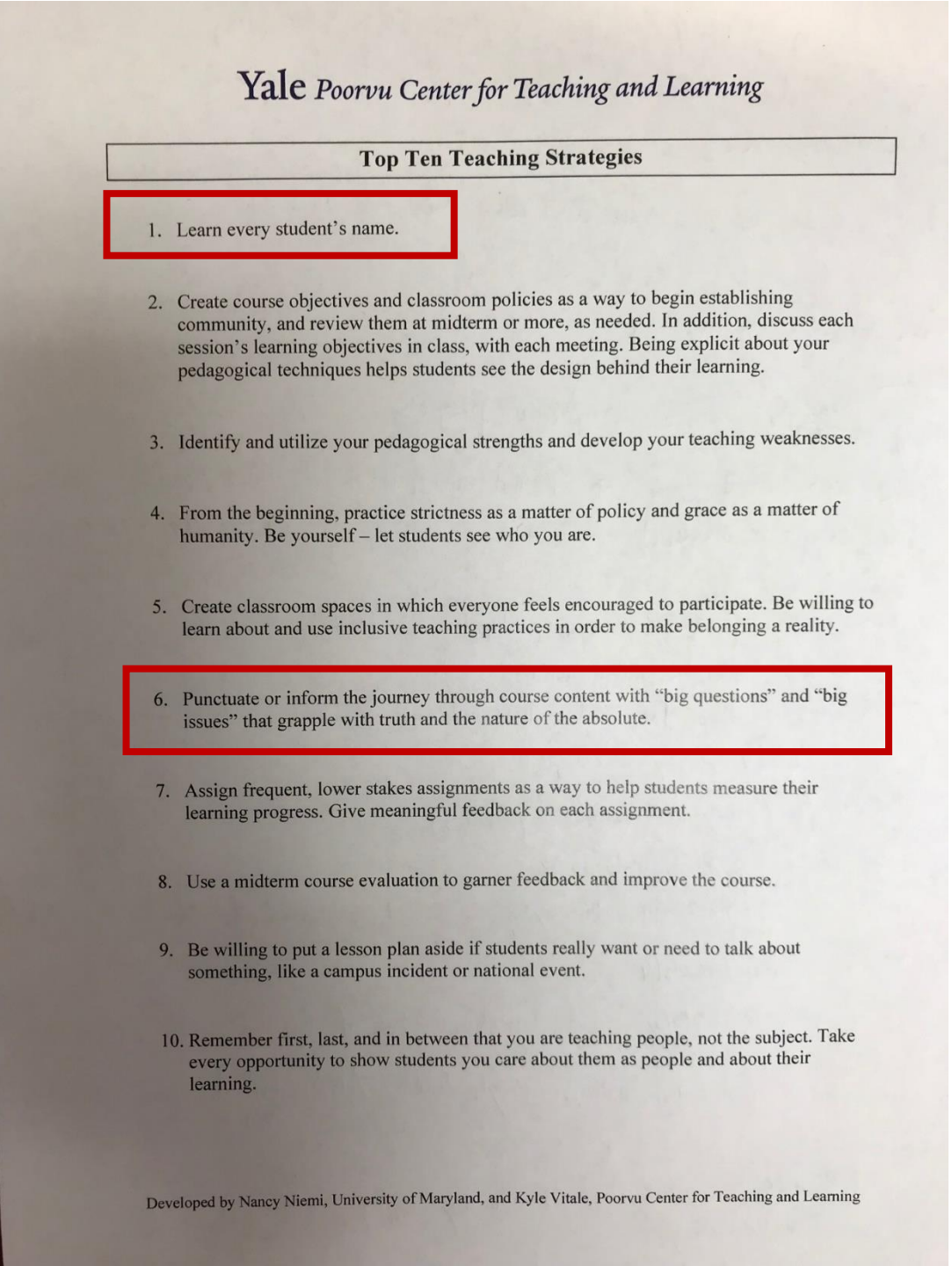
"Boy, those French: They have a different word for everything!"

- Steve Martin

Boy, those Statisticians: They use common words to mean something different!

Bias,
confidence,
significance

Central concepts in 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

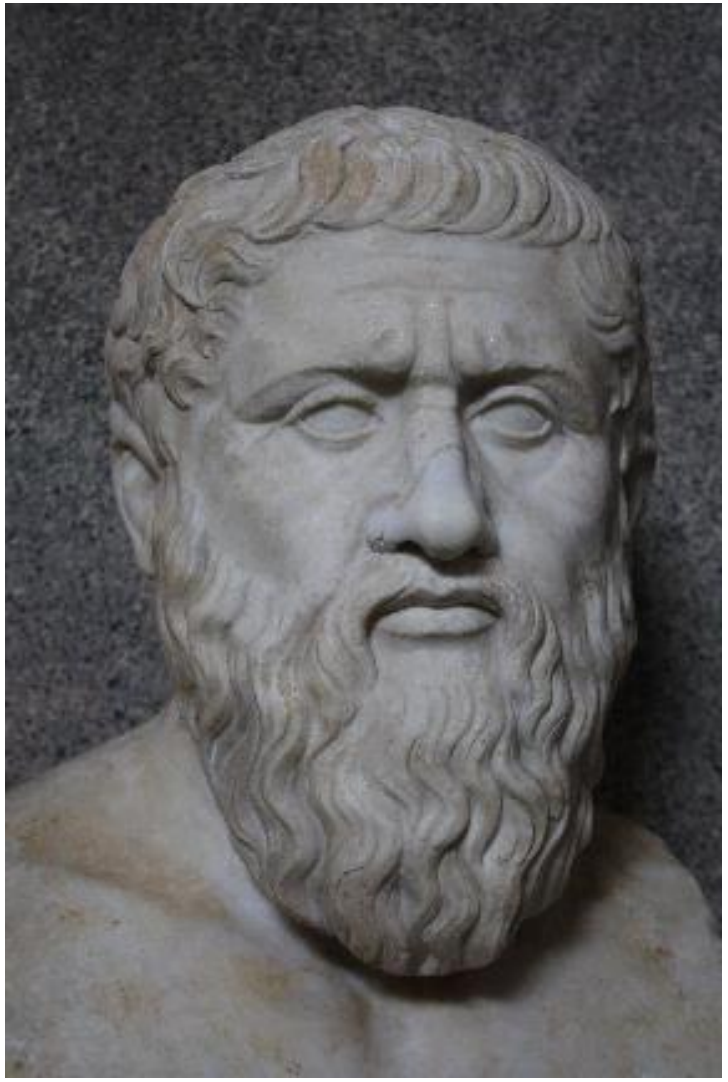
Center for Teaching and Learning tips

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

Central concepts in Statistics

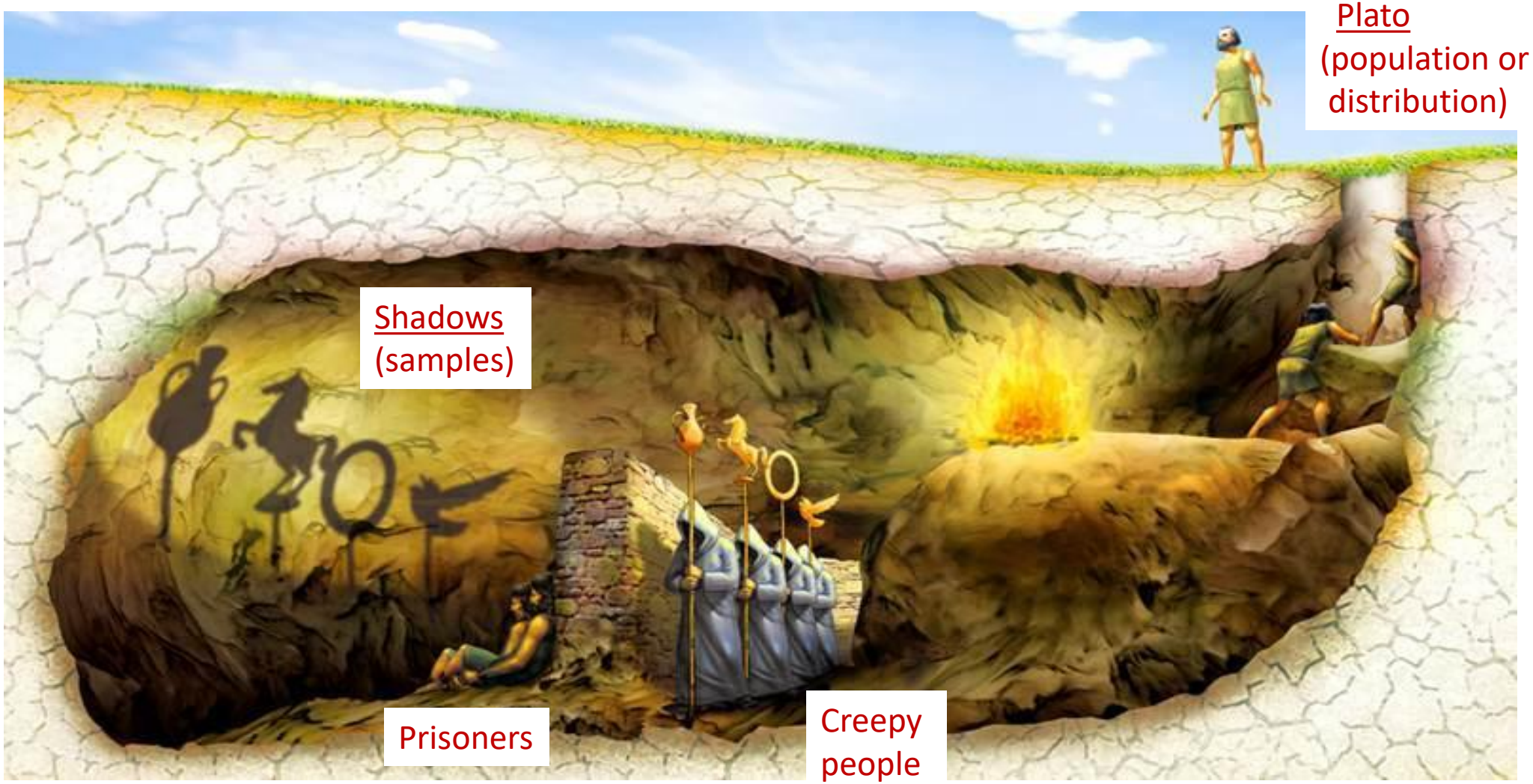
The Truth!



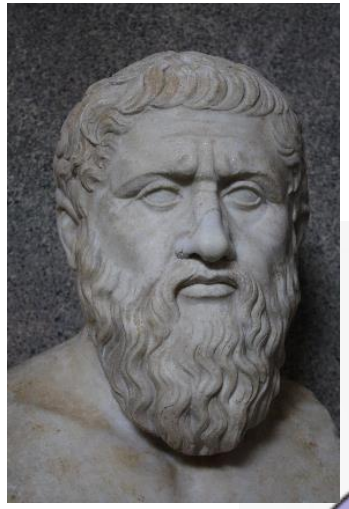
If we could see all the (infinite) data, we would know the Truth®!

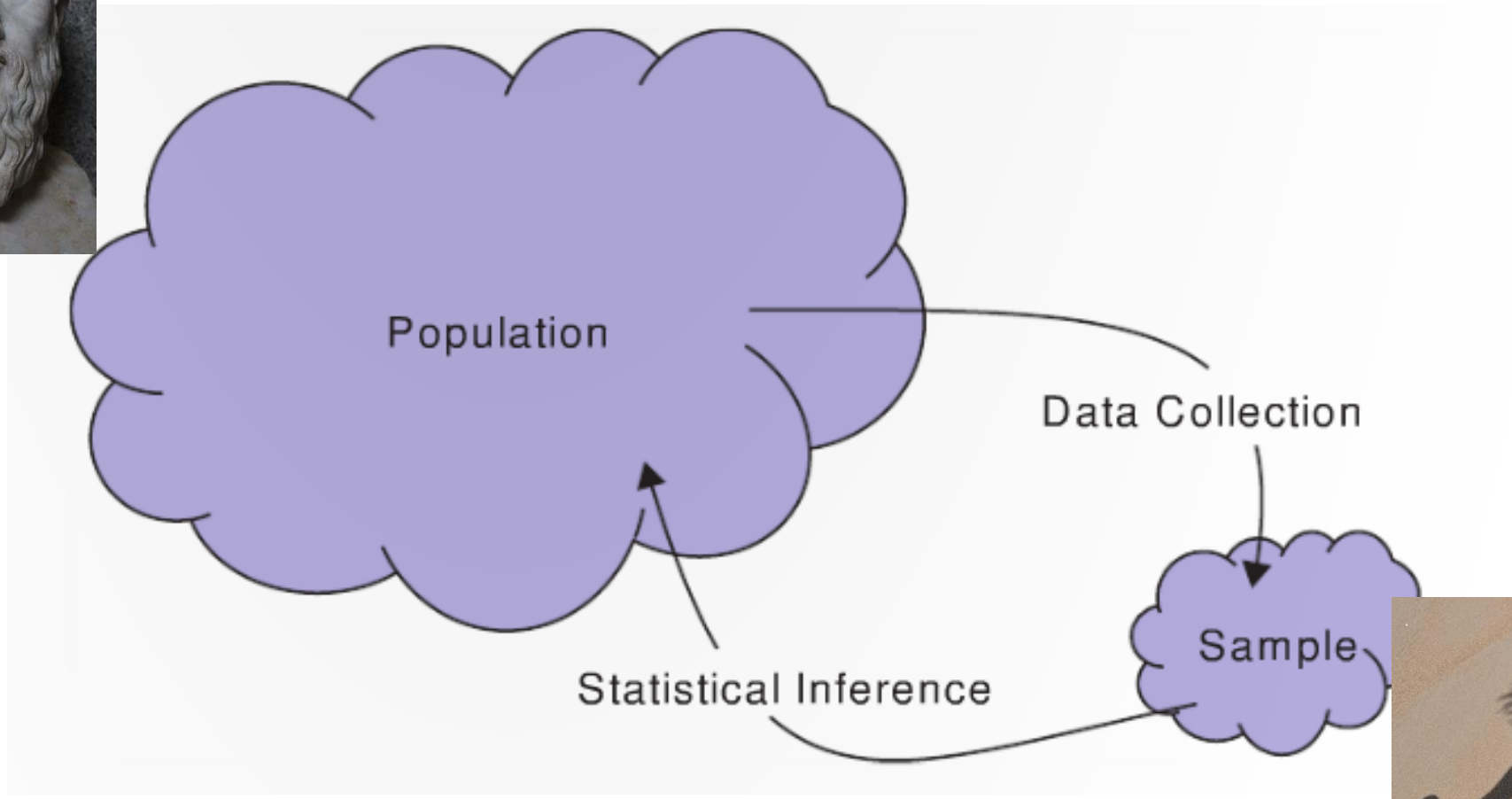
Alas, we can only see a small subset of the data (a sample) so we merely see a shadow of the truth

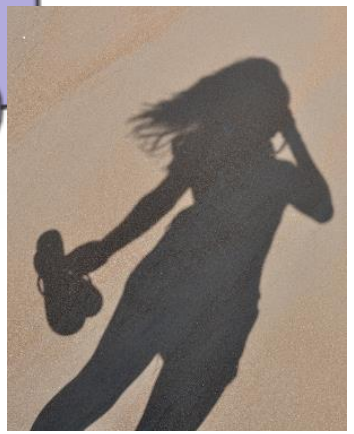
Plato's cave



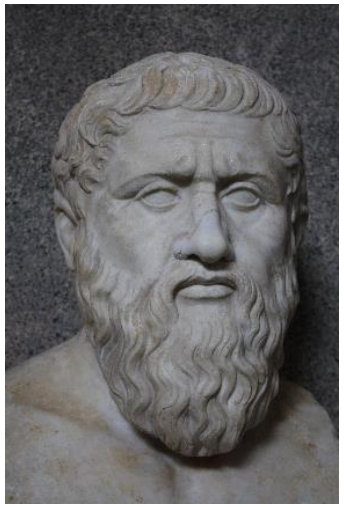
From The Republic (~ 380 BCE)




Population
Data Collection
Sample
Statistical Inference



Sample from a Population



Population: all individuals/objects of interest

Sample: A subset of the population

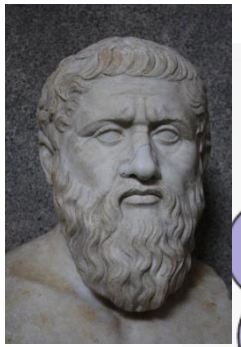
Descriptive and inferential statistics

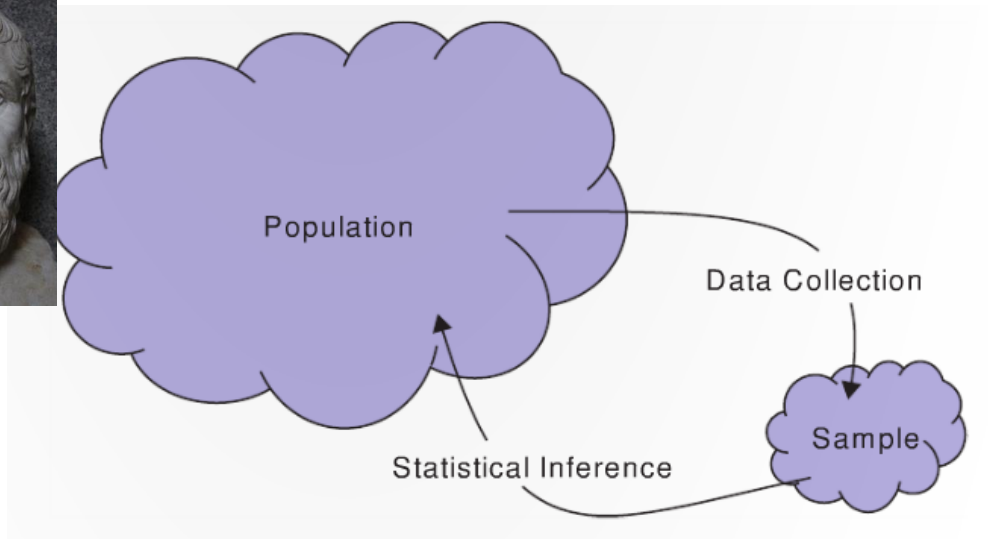
Descriptive Statistics: describe the sample of data we have

- i.e., describe the shadows

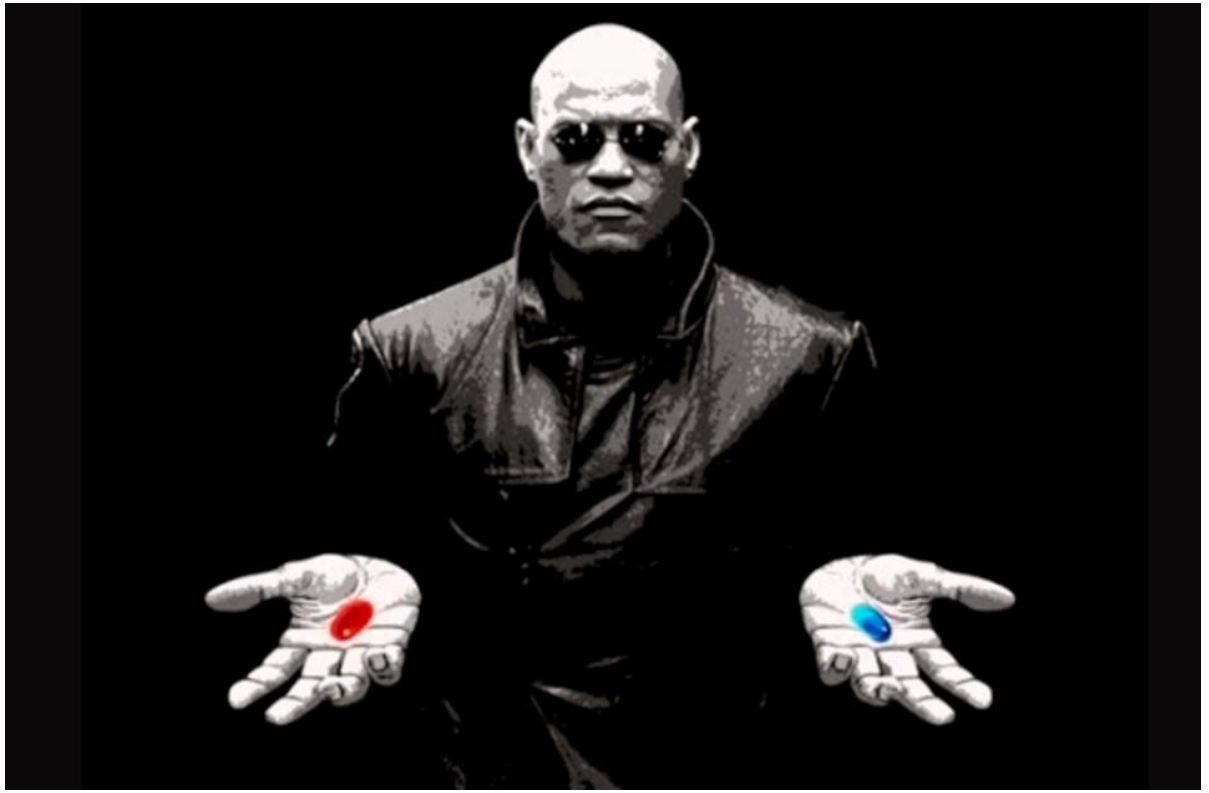
Inferential Statistics: use the sample to make claims about properties of the population/process

- i.e., try to use the data to get at the truth





Can you handle the Truth?



If not, perhaps you should not take this class

You've been warned…

Structured data: exploring the shadows

An Example Dataset (Shadows)

Variables

Cases

	Year	Smoke	Award	HigherSAT	Exercise	TV	Height
1	Senior	No	Olympic	Math	10.0	1	71
2	Sophomore	Yes	Academy	Math	4.0	7	66
3	FirstYear	No	Nobel	Math	14.0	5	72
4	Junior	No	Nobel	Math	3.0	1	63
5	Sophomore	No	Nobel	Verbal	3.0	3	65
6	Sophomore	No	Nobel	Verbal	5.0	4	65
7	FirstYear	No	Olympic	Math	10.0	10	66
8	Sophomore	No	Olympic	Math	13.0	8	74

An Example Dataset (Shadows)

Categorical Variable

Quantitative Variable

Cases (observational units)

	Year	Smoke	Award	HigherSAT	Exercise	TV	Height
1	Senior	No	Olympic	Math	10.0	1	71
2	Sophomore	Yes	Academy	Math	4.0	7	66
3	FirstYear	No	Nobel	Math	14.0	5	72
4	Junior	No	Nobel	Math	3.0	1	63
5	Sophomore	No	Nobel	Verbal	3.0	3	65
6	Sophomore	No	Nobel	Verbal	5.0	4	65
7	FirstYear	No	Olympic	Math	10.0	10	66
8	Sophomore	No	Olympic	Math	13.0	8	74

Edmunds transaction data

- What are the observational units (cases)?
- Which variables are: quantitative or categorical?

	transactionid	date_sold	make_bought	price_bought	zip_bought	mileage_bought	color_bought
1	16966151	2014-09-27	Acura	30892.00	21043	40	BLACK
2	16914863	2014-09-27	Toyota	25566.00	15108	297	SILVER
3	15977620	2014-07-31	Nissan	34300.00	8753	0	JAVA
4	18666685	2015-01-27	Subaru	30059.00	7446	10	CRYSTAL WHITE PEARL
5	14383133	2014-04-27	Honda	32508.00	97027	21	MODERN STEEL
6	18196788	2014-12-18	Toyota	10819.66	95117	55246	WHITE
7	15722278	2014-07-24	Audi	59630.00	90401	143	GLACIER WHITE

Summary of concepts

1. Population: all individuals/objects of interest (Truth)

2. Sample: A subset of the population (shadows)

3. Statistical inference: Making judgments about the population using data from the sample

4. Structured data has

- Cases/observational units: rows in a data set
- Variables: columns in a data set

5. Variables can be

- Categorical: fall into discrete categories
- Quantitative: are numbers

For next class

Please fill out the class survey on Canvas under quiz

Practice problems from Lock 5, first edition:

1.1, 1.3, 1.5, 1.11, 1.25, 1.26

Chapter 1 is posted on Canvas