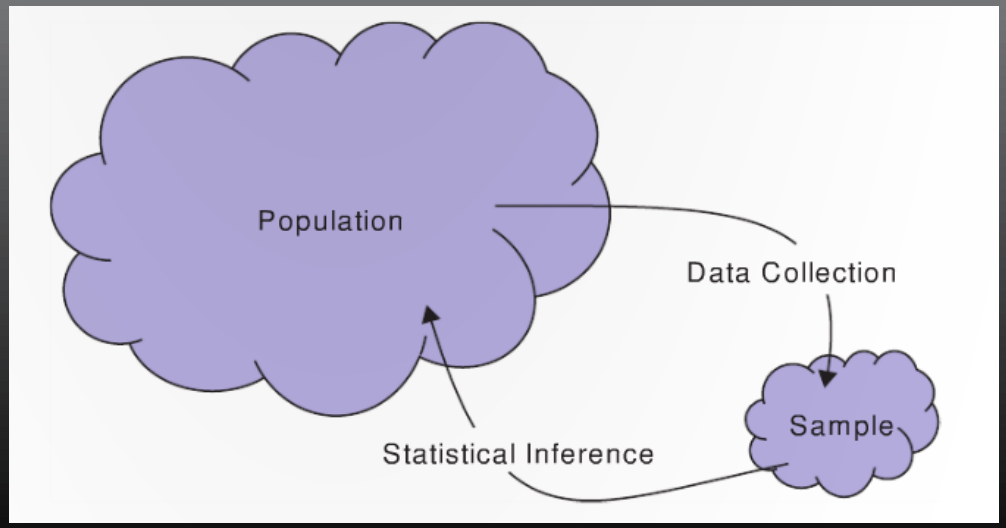
Class 1: logistics and central concepts in Statistics



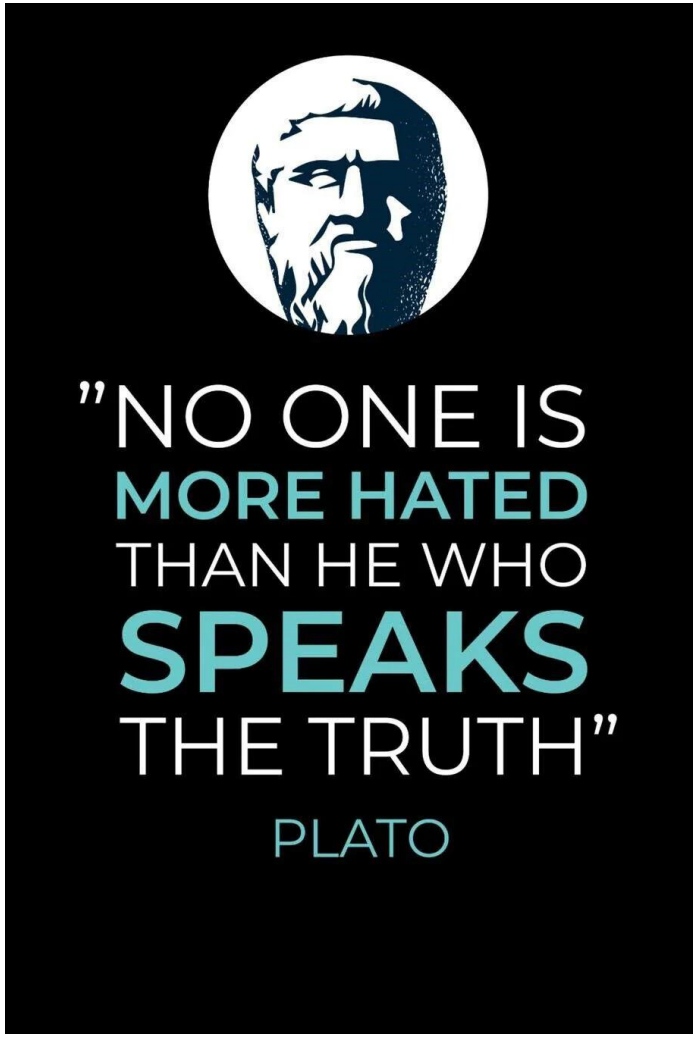
Overview

Class logistics

What is Statistics?

Central concepts in Statistics

Structured data



Course overview and logistics

Office hours and contact information

Email: ethan.meyers@yale.edu

Office hours: 2-3pm Mondays and Wednesdays

Office: Kline Tower room 1253

- https://yale.zoom.us/j/93177892725

Teaching Staff

Preceptors

- Lynda Aouar: lynda.aouar@yale.edu
- Addison McGhee: addison.mcghee@yale.edu

Course Manager

- Brian Xiang: brian.xiang@yale.edu

Teaching Fellow

- Zhenia Rudyk: zhenia.rudyk@yale.edu
- Vladimir Averin: vladimir.averin@yale.edu
- Steven Ward: steven.ward@yale.edu
- Ben Green: beniamino.green@yale.edu

Undergraduate Learning Assistants

- Cindy Cai: c.cai@yale.edu
- Kathleen Cain: kathleen.cain@yale.edu
- Alyssa Chang: alyssa.chang@yale.edu
- Jessica Huang: jessica.huang.jh3359@yale.edu
- Eric Lin: eric.lin.el832@yale.edu
- Jasmine Garcia: jasmine.garcia@yale.edu
- Asher Mehr: asher.mehr@yale.edu
- Sarah Lepkowitz: sarah.lepkowitz@yale.edu
- Lucas Papamitsakis: lucas.papamitsakis@yale.edu
- Aryav Bothra: aryav.bothra@yale.edu

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

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!

Plan for the semester

Exploring data/descriptive statistics (weeks 1-4)

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.

Practice sessions

Addison and Lynda will be hosting one-hour practice sessions each week

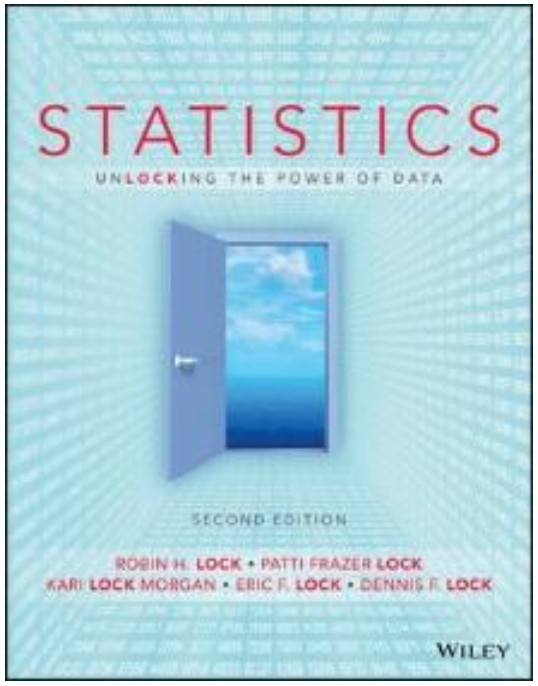
- Each session will be offered at several different times each week
- Please fill in class survey to let us know what times work for you

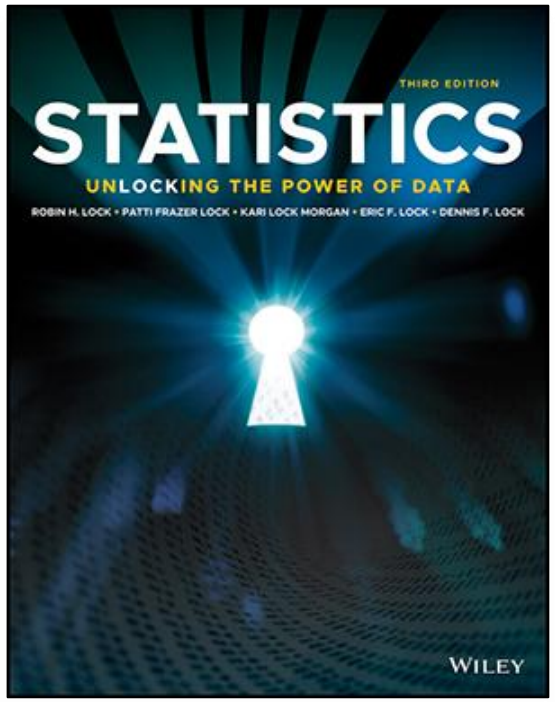
The practice sessions are a great opportunity to deepen your understanding of statistics concepts, practice analyze data using R, and to get your questions answered!

Highly recommended to attend these sessions

- Attendance is optional
- If you score below a median cutoff score on the exams, a point will be added to your score for each weekly practice session you attend (up to 5 points)

Textbook: Lock5





Additional reading and other resources will be posted to Canvas: https://yale.instructure.com/courses/109046

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 is available by Tuesday's class and is due at 11pm on Sundays
- Late homework (90%) credit if turned in by 11pm on Monday
  - For any other extensions a Deans Excuse is needed
- Lowest scoring worksheet will be dropped

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.

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: October 23rd during the regular class time (15%)
   - Final: December 15th at 2pm (28%)

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

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 (~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!

Accommodations and Academic honesty

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

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

ChatGPT and other LLMs

You can use LLMs 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 into chatGPT

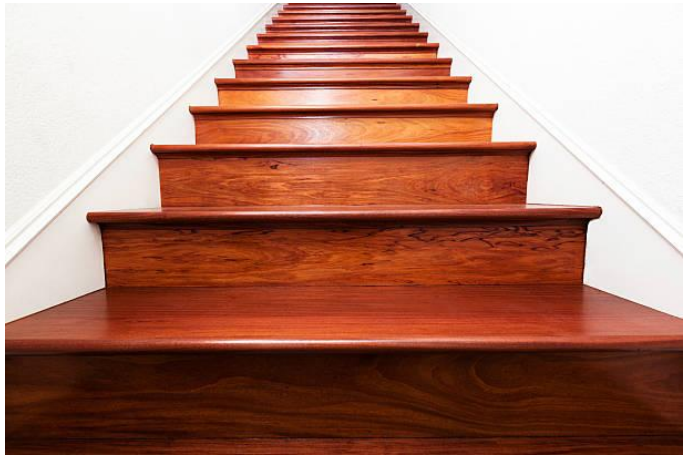
If it appears your homework answers were generated by ChatGPT or another LLM (Claude, Gemini, etc.) you will be Yale Executive Committee

- Also, if you cheat on the homework you will do very poorly on the exams

How to be successful in this class

Keep up with the work!

- The class might seem easy at times, but if you skip steps you will be in trouble



My role is coach

- I try to make the material easy to understand and engaging
  - Regular -> flipped -> regular

- I can’t workout for you

Class surveys

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

- Under the Quizzes link on the left

Also, fill out the practice session survey to let Lynda and Addison know what days/times work for having practice sessions

Any questions about the class logistics???

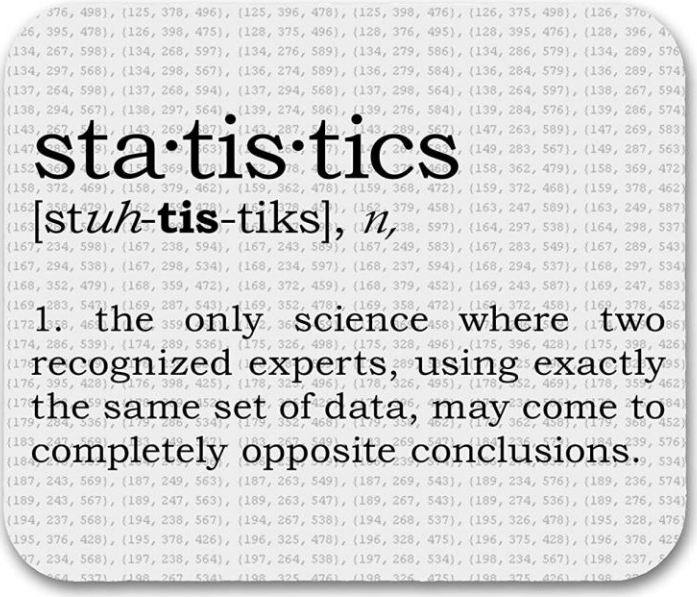
- Ask on Ed Discussions!

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)

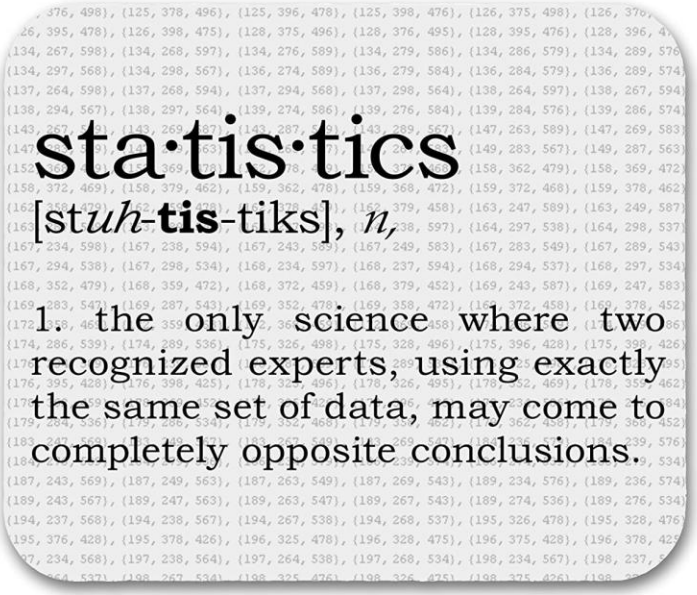


Fienberg, S. (2014). What is Statistics? The *Annual Review of Statistics and Its Application,* 1:1-9

My thoughts

Statistics develops methods concerning how 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
  - Often the methods reply on probability models



Statistical analyses are <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

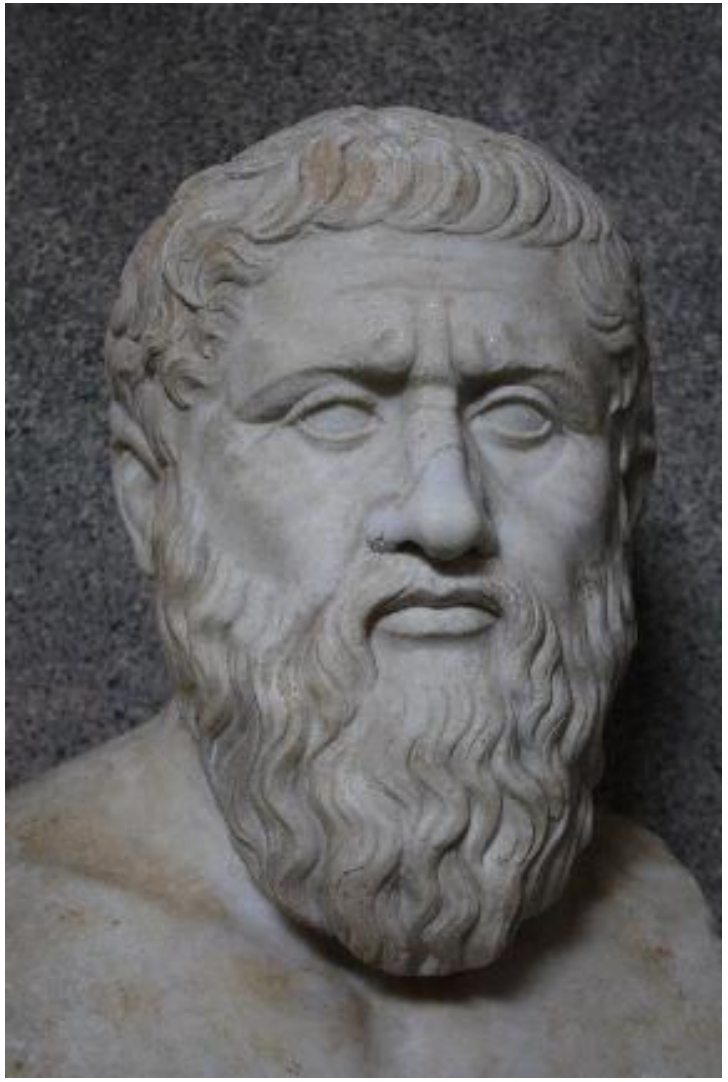
Boy, those Statisticians: They use common words to mean something different!

Bias,
confidence,
significance

Central concepts in Statistics

Central concepts in Statistics

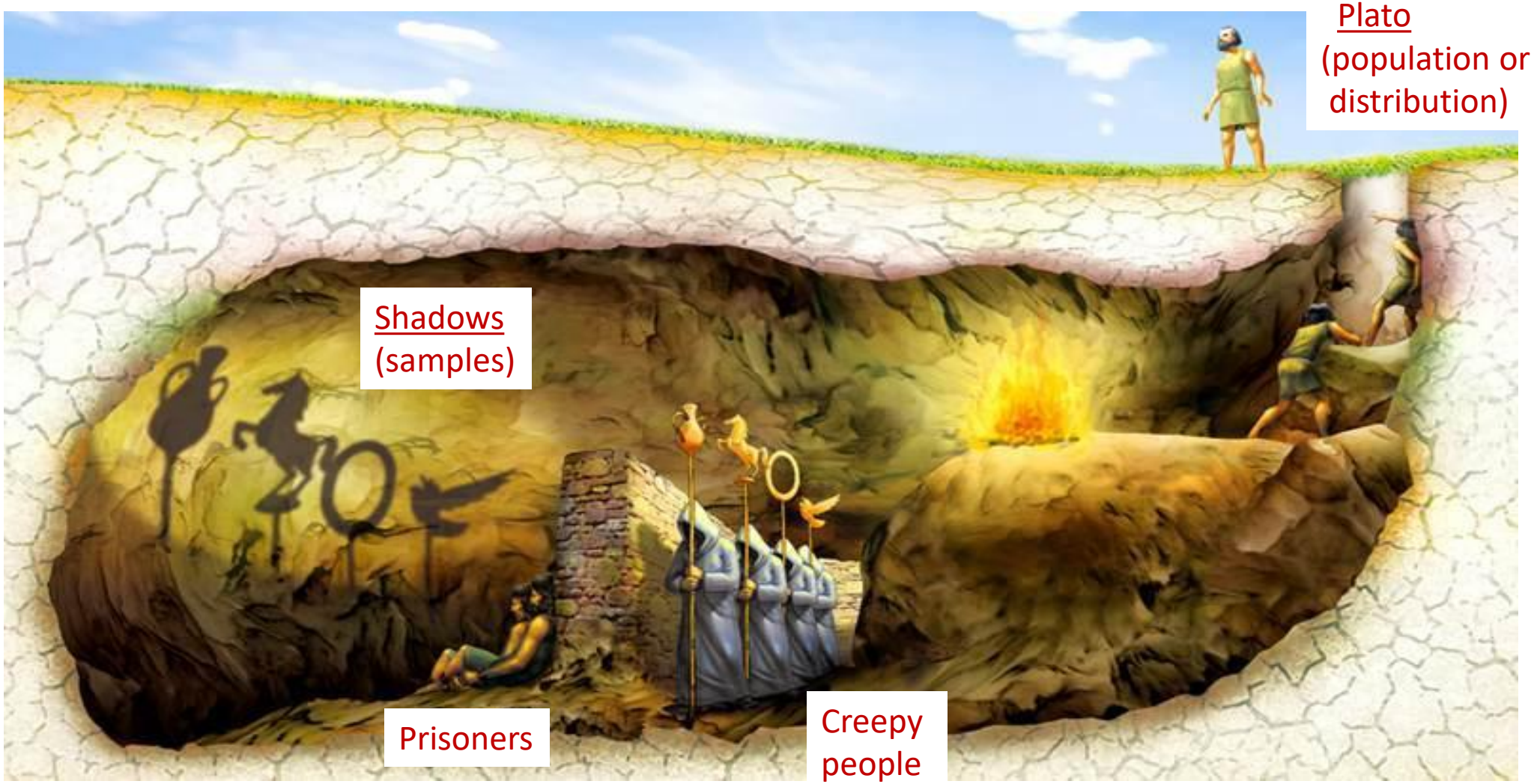
The Truth!



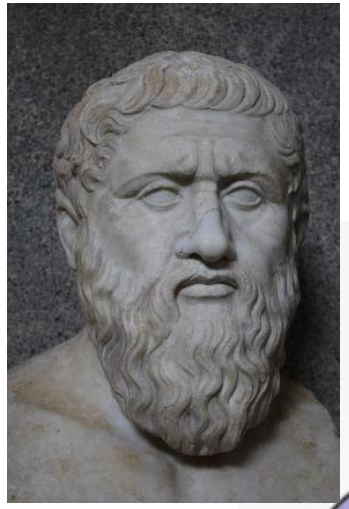
If we could see all the (infinite) data, we would know the Truth®!

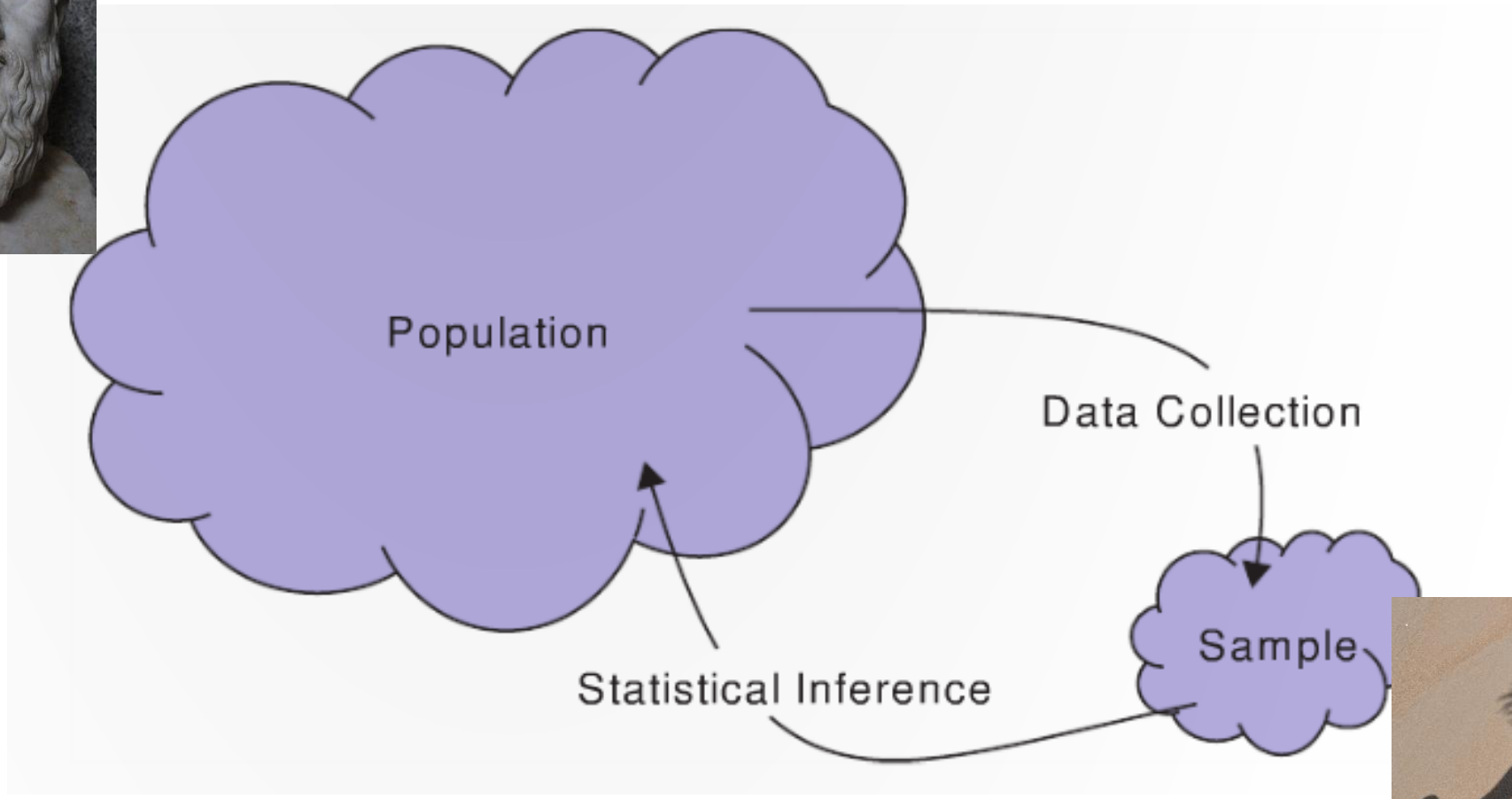
Alas, we can only see a small subset of the data (a sample) so we merely see a shadow of the truth

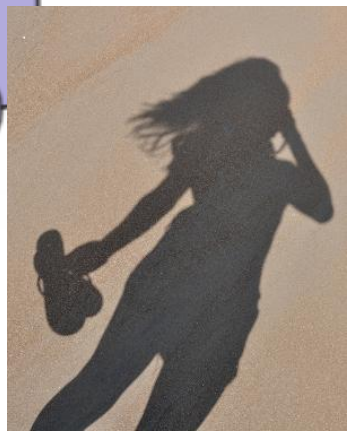
Plato's cave



From The Republic (~ 380 BCE)

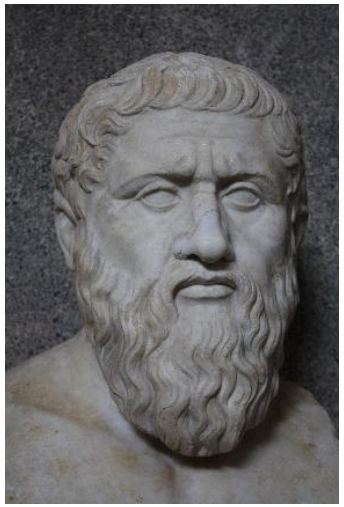



Population
Data Collection
Sample
Statistical Inference



Sample from a Population

Population: all individuals/objects of interest



Sample: A subset of the population

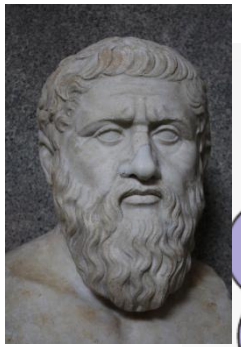
Descriptive and inferential statistics

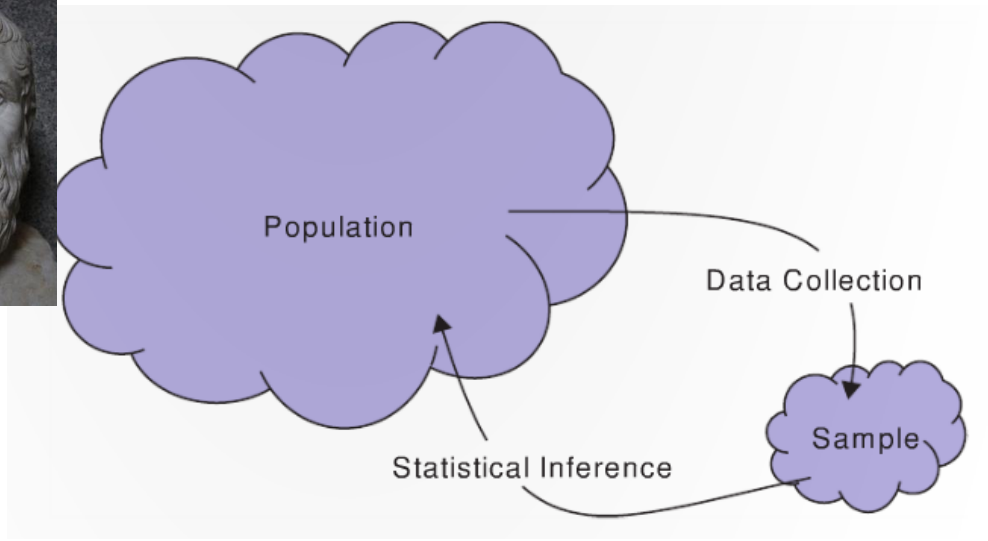
Descriptive Statistics: describe the sample of data we have

- 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 practice session, and background surveys on Canvas under Quizzes

Try some practice problems from Lock 5 textbook

- Section 1.1, first half of section 1.2

Try to connect to the class RStudio server