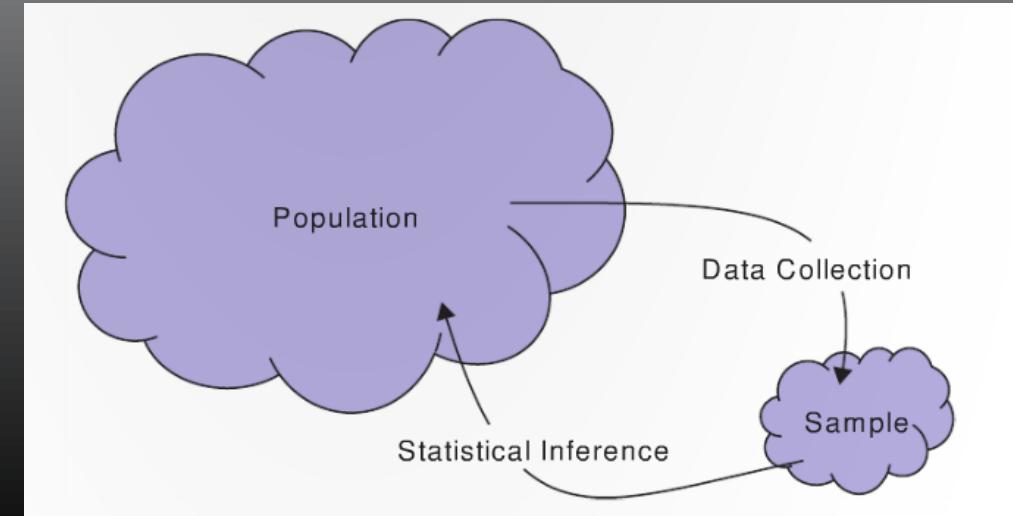


# Class 1: logistics and central concepts in Statistics



# Overview

Class logistics

What is Statistics?

Central concepts in Statistics

Structured data

If there is time: Quarto!

# Course overview and logistics



# Office hours and contact information

Email: [ethan.meyers@yale.edu](mailto:ethan.meyers@yale.edu)

Office hours: 2-3pm Wednesdays and Fridays

Office: Kline Tower room 1253

- Also on zoom: <https://yale.zoom.us/j/98789816394>

# Teaching Staff



## Preceptors

- Lynda Aouar: [lynda.aouar@yale.edu](mailto:lynda.aouar@yale.edu)
- Addison McGhee: [addison.mcghee@yale.edu](mailto:addison.mcghee@yale.edu)

## Course Manager

- Kexin Shi : [kexin.shi@yale.edu](mailto:kexin.shi@yale.edu)

## Teaching Fellow

- Natalie Hernandez : [natalie.hernandez@yale.edu](mailto:natalie.hernandez@yale.edu)
- Lianhan Shang: [lianhan.shang@yale.edu](mailto:lianhan.shang@yale.edu)
- Brandon Tran: [brandon.tran@yale.edu](mailto:brandon.tran@yale.edu)
- Liya Xiang: [liya.xiang@yale.edu](mailto:liya.xiang@yale.edu)
- Ruth Zheng: [ruth.zheng@yale.edu](mailto:ruth.zheng@yale.edu)

## Undergraduate Learning Assistants

- Cindy Cai: [c.cai@yale.edu](mailto:c.cai@yale.edu)
- Leya Dereje: [leya.dereje@yale.edu](mailto:leya.dereje@yale.edu)
- Mia Gawith: [mia.gawith@yale.edu](mailto:mia.gawith@yale.edu)
- Jessica Huang: [jessica.huang.jh3359@yale.edu](mailto:jessica.huang.jh3359@yale.edu)
- Eric Lin: [eric.lin.el832@yale.edu](mailto:eric.lin.el832@yale.edu)
- Jasmine Garcia: [jasmine.garcia@yale.edu](mailto:jasmine.garcia@yale.edu)
- Asher Mehr: [asher.mehr@yale.edu](mailto:asher.mehr@yale.edu)
- Sarah Lepkowitz: [sarah.lepkowitz@yale.edu](mailto:sarah.lepkowitz@yale.edu)
- Cleia Jones: [cleia.jones@yale.edu](mailto:cleia.jones@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

- Descriptive statistics and plots
- Inferential statistics:
  - Hypothesis tests and confidence intervals
  - Using randomization and parametric methods

## 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

February 5<sup>th</sup>

Relationships between variables

- Correlation and regression



# Plan for the semester

## Inferential Statistics using computational methods

Sampling distributions

Confidence intervals

- The bootstrap

Hypothesis tests using randomization methods

- Permutation tests

March 5<sup>th</sup>

## Inferential Statistics using mathematical models

Hypothesis tests and confidence intervals using parametric methods

- T-tests, ANOVA, etc.



# Practice sessions



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

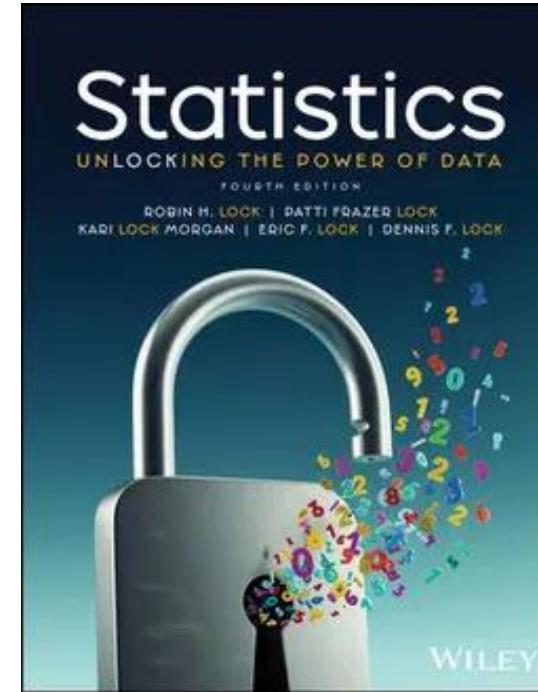
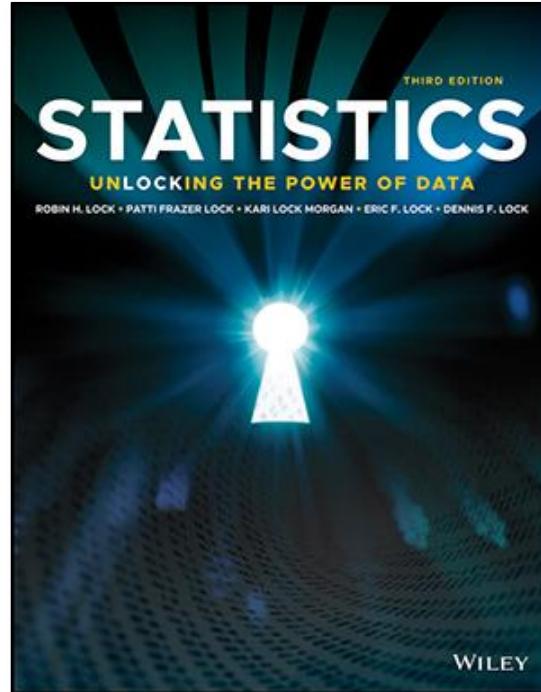
- Each session will be offered at several different times each week
- Please fill in class survey to let us know what times work for you

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** (although 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/115642>

# Assignments and grades

## 1. Homework problem sets (36%)

- 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
  - (with a 59 minute grace period)
- 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:** Can dolphins communicate abstract ideas?



# 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 (52% total)

- Mini-exam: February 5<sup>th</sup> during regular class time (6%)
- Midterm: March 5<sup>th</sup> during the regular class time (16%)
- Final: May 5<sup>th</sup> at 7pm (30%)

## 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 lab/homework questions**

- i.e., Do not type a lab/homework question into ChatGPT
- Do not have it complete code you started writing (or cut and paste any code from an LLM)
- You are responsible for understanding all code that appears on your homework solutions

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

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

# How to be successful in this class

To do well in this class you need to keep up with the work!

- The class might seem easy at times, **but if you fall behind you will be in trouble**

My role is to provide resource and be a coach

- I try to make the material easy to understand and engaging
  - But I can't do the workout for you
  - (nor can I change your grade after the class)
- **Attend the practice sessions!!!**

# Class surveys

In order to get to know you and to adjust the class to everyone's interests, please:

1. Fill out the background survey on canvas
2. 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 rely on probability models

Statistical analyses are part 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

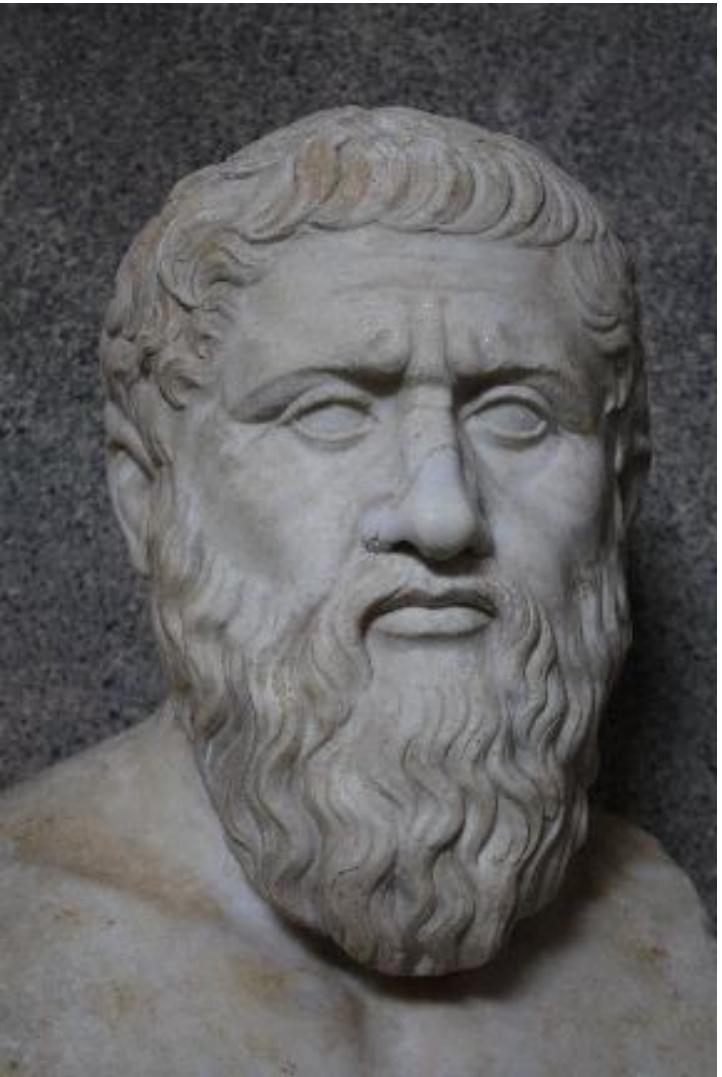
# Central concepts in Statistics

# Central concepts in Statistics



THE TRUTH IS OUT THERE

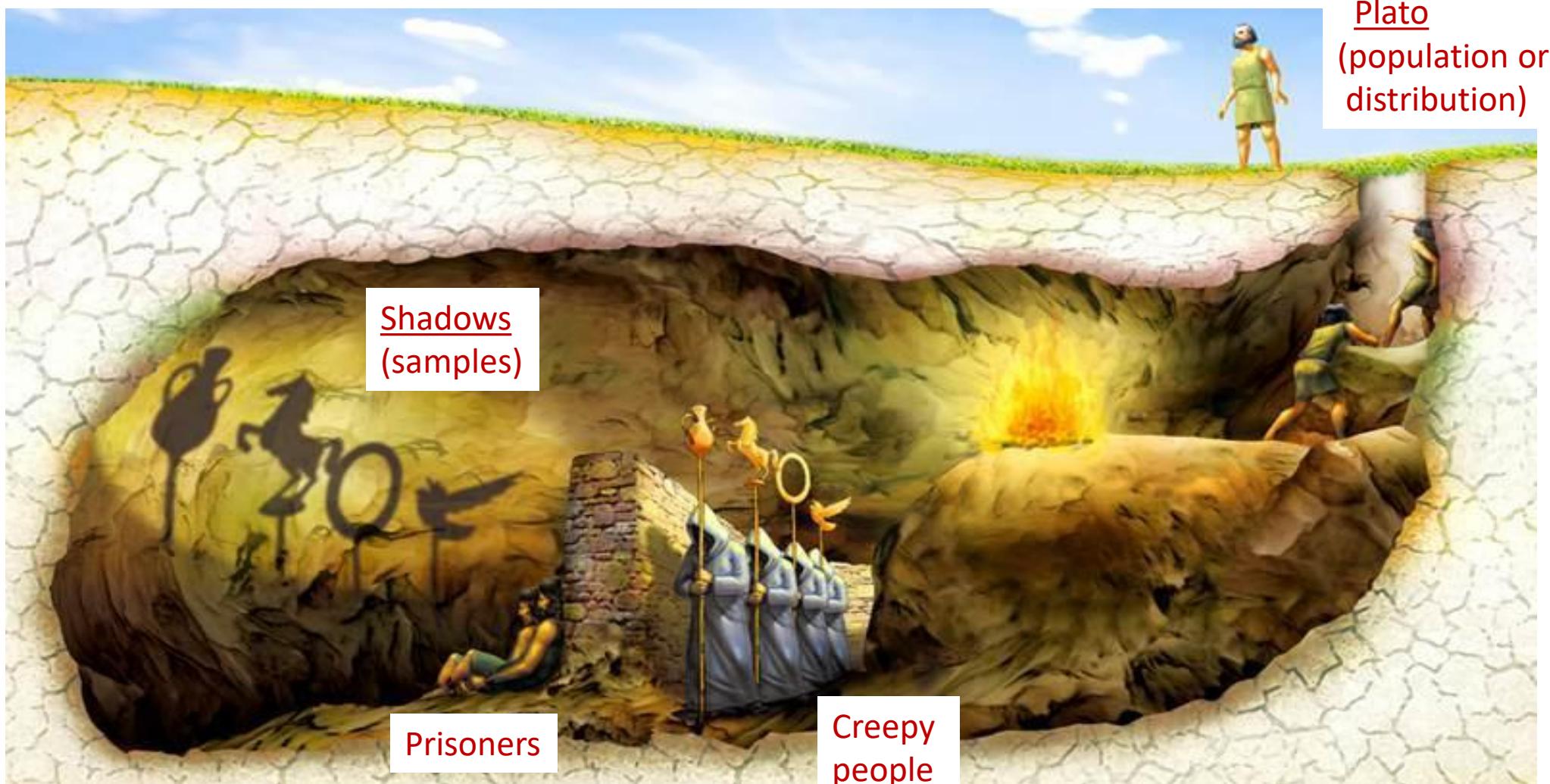
# The Truth!



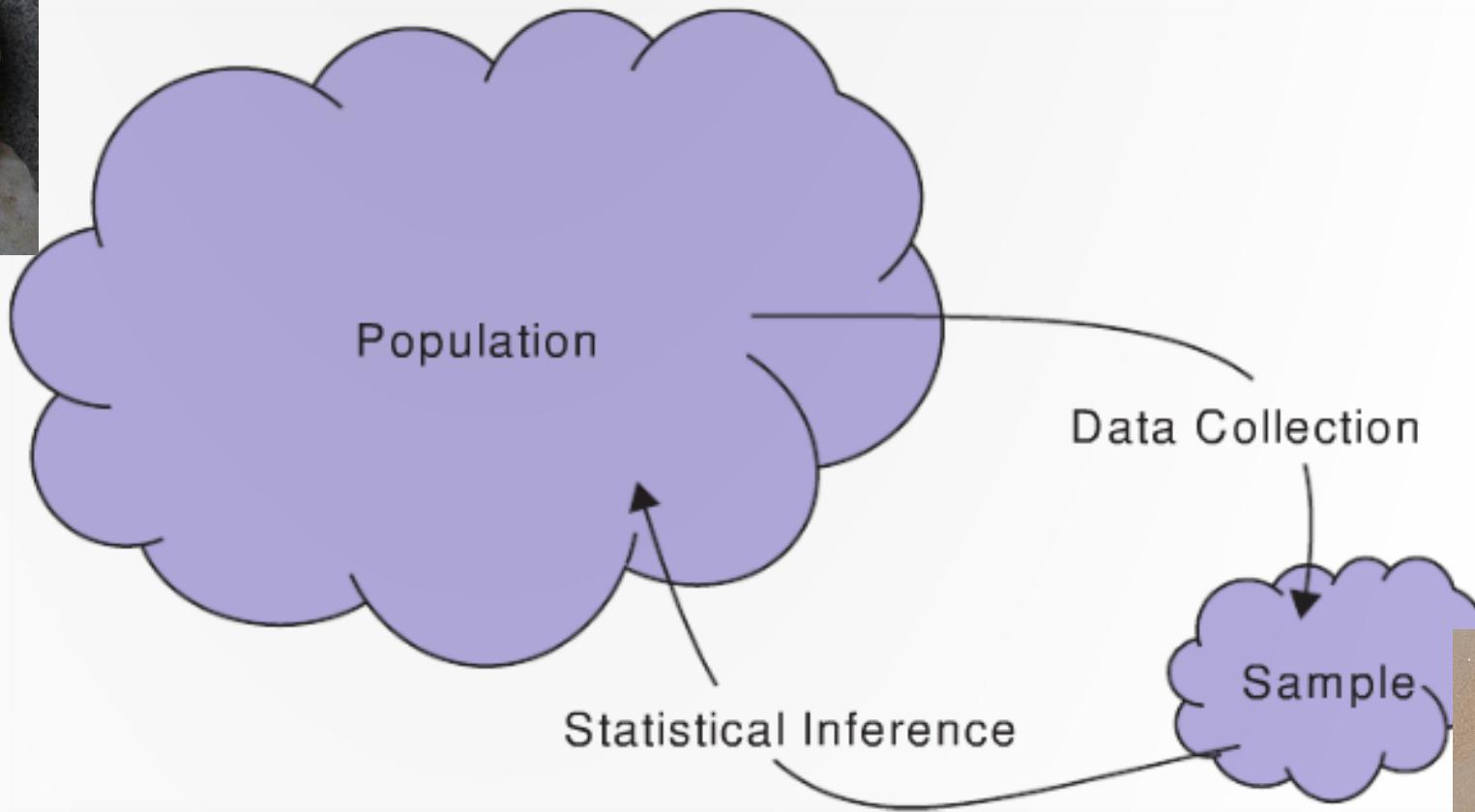
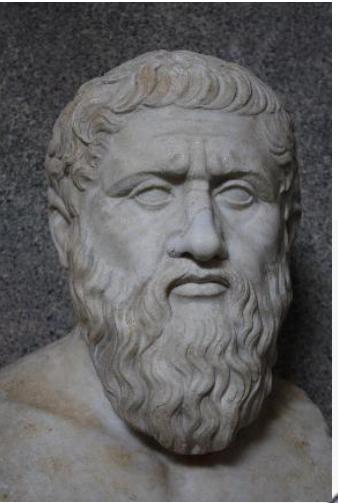
If we could see all the (infinite) data, we would know the Truth®!

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

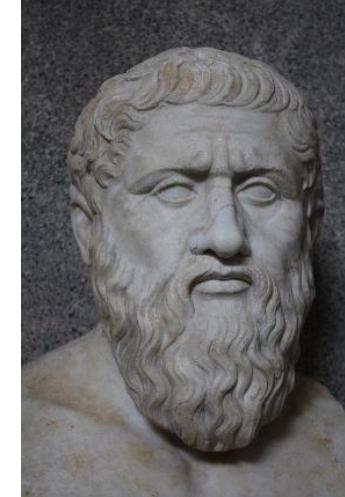


From The Republic (~ 380 BCE)



# Sample from a Population

**Population:** all individuals/objects of interest



**Sample:** A subset of the population



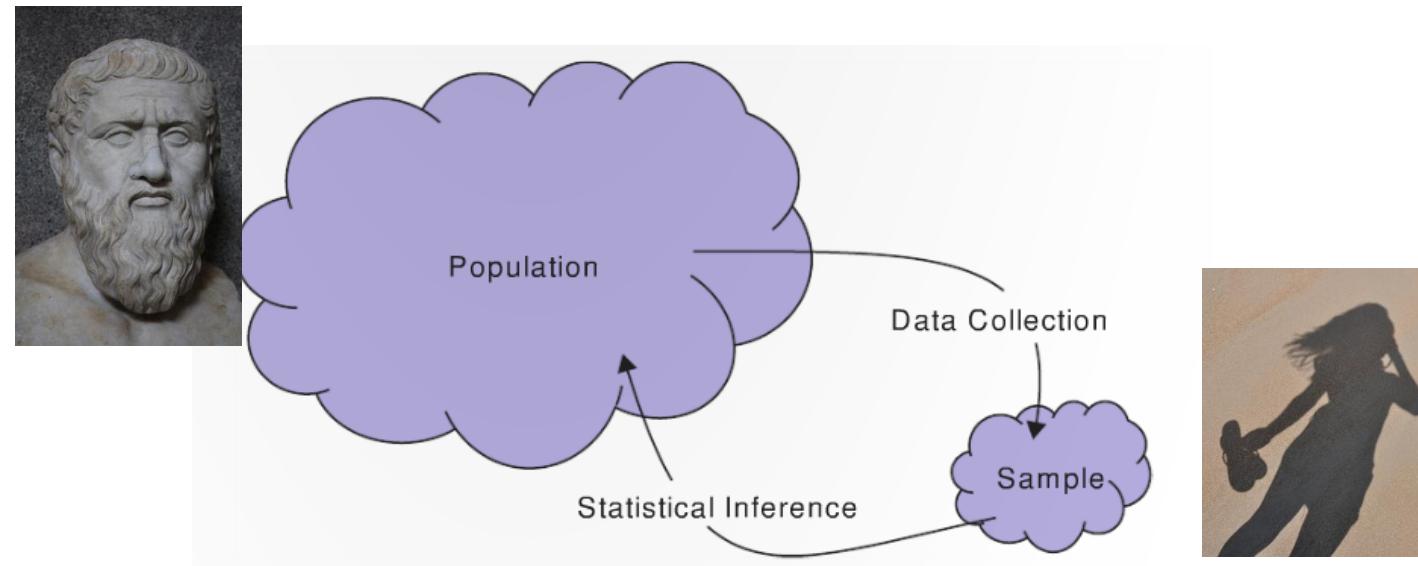
# Descriptive and inferential statistics

**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



# Structured data: exploring the shadows



# An Example Dataset (Shadows)

## Variables

Cases

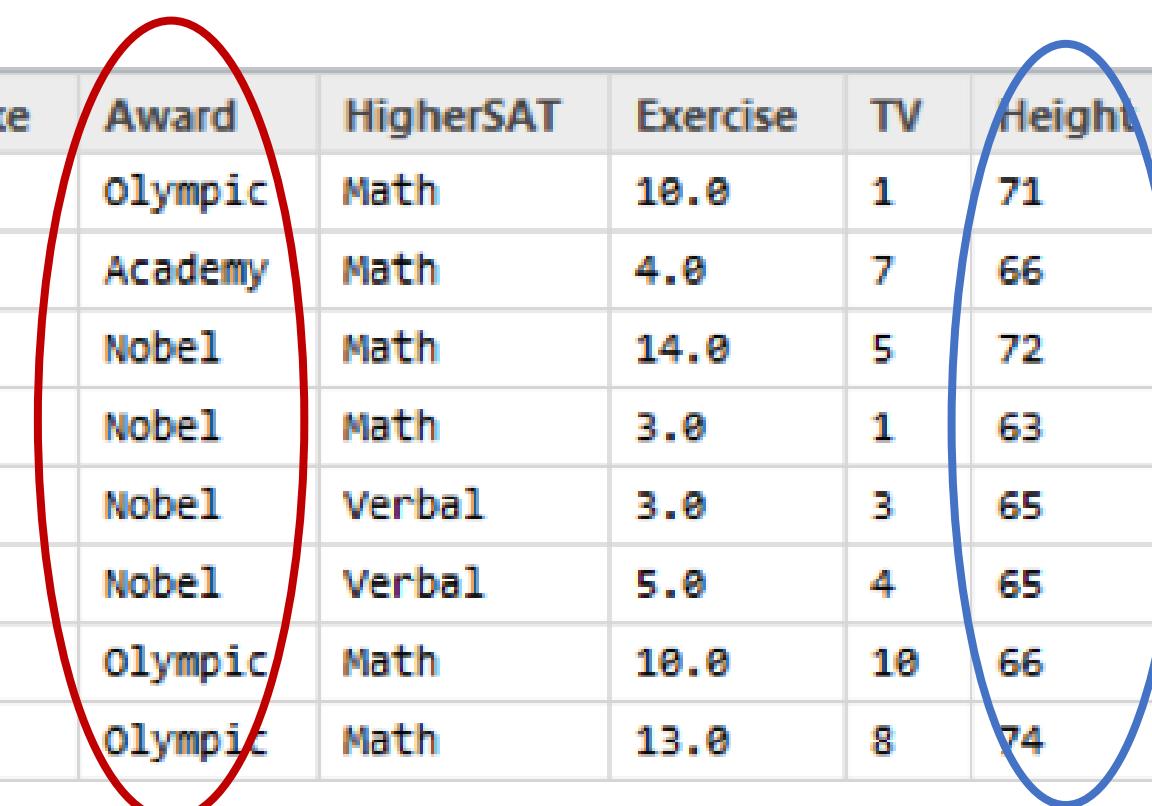
	Year	Smoke	Award	HigherSAT	Exercise	TV	Height
1	Senior	No	Olympic	Math	10.0	1	71
2	Sophomore	Yes	Academy	Math	4.0	7	66
3	FirstYear	No	Nobel	Math	14.0	5	72
4	Junior	No	Nobel	Math	3.0	1	63
5	Sophomore	No	Nobel	Verbal	3.0	3	65
6	Sophomore	No	Nobel	Verbal	5.0	4	65
7	FirstYear	No	Olympic	Math	18.0	10	66
8	Sophomore	No	Olympic	Math	13.0	8	74

# An Example Dataset (Shadows)

Cases  
(observational units)

Categorical Variable      Quantitative Variable

	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	18.0	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



Quarto

# Quarto

Quarto (.qmd files) allow you to embed written descriptions, R code and the output of that code into a nice looking document

Creates a way to do reproducible research!



# Quarto

Everything in R chunks is executed as code:

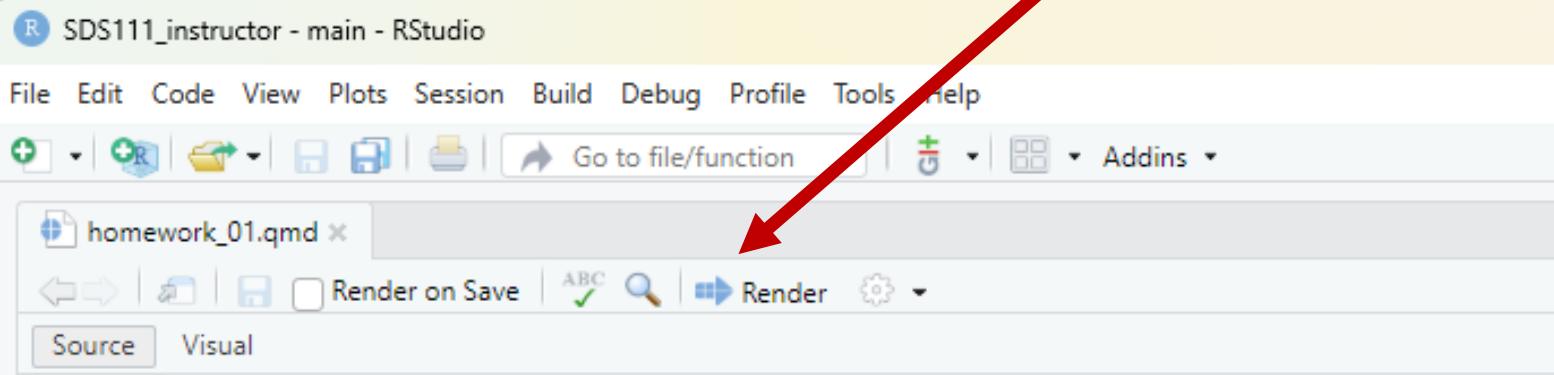
```
```{r}
  # this is a comment
  # the following code will be executed
  2 + 3
```

```

Everything outside R chunks appears as text

# Render to a pdf

Turns a Quarto document to a pdf



```
1 ---  
2 title: "Homework 1"  
3 format: pdf  
4 editor: source  
5 editor_options:  
6   chunk_output_type: console  
7 ---  
8  
9  
10  
11  
12 $\\$  
13  
14  
15  
16  
17 # Welcome to the first homework assignment!  
18
```

# Formatting in Quarto

We can add formatting to text outside the code chunks

Examples:

## Level 2 header

**\*\*bold\*\***

LaTeX {  
    \$ \pi \$  
    \$ x\_{\text{outcome}} \$}

# Avoid hard to debug code!

Only change a few lines at a time and then render your document to make sure everything is working!

**I.e., render your to pdf document often!**



# Questions?



Let's explore Quarto by connecting to the Rstudio server!

For next class



Please fill out:

1. Background survey
2. Practice session survey

Try some practice problems from Lock 5 textbook

- Section 1.1, first half of section 1.2