

CS 206

Introduction to Statistics

Ethan Meyers

Overview

Introductions

Go over the syllabus

What is statistics?

Samples and Populations (shadows and truth)

Structured data: quantitative and categorical variables

R Studio

Introductions

About you:

- Name
- Preferred gender pronouns
- Where you are coming from
- Year at Hampshire (Div I, II, III)
- Anything else you want to share

Why are you interested in this class?

General Information

My contact info:

Email: emeyers@hampshire.edu

Office: ASH 133

Office hours (ASH 133) M 2:30-3:30, W 11-12

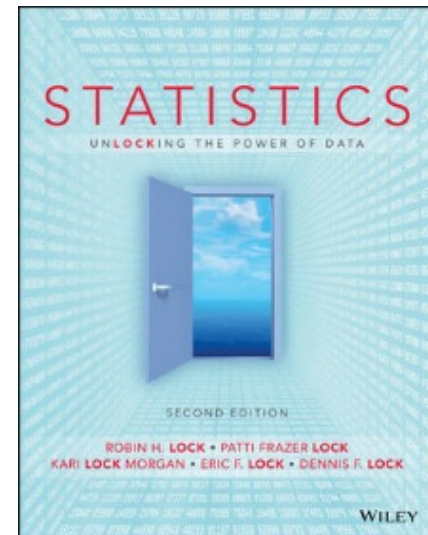
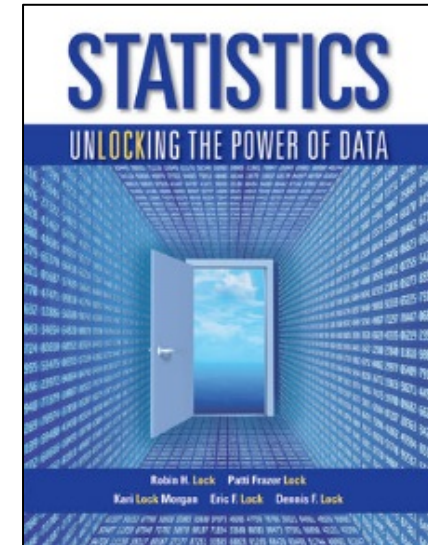
Email me for additional hours

Course objectives:

- Understand the key concepts in Statistics
- Learn how to analyze real data (R)

Textbook: Lock5

Website: <https://moodle.hampshire.edu/course/view.php?id=6503>



Assignments

1. Readings/practice problems from Lock 5 textbook

- This readings will supplement what is discussed in class
- Practice problems will make sure you understand the concepts (not collected)

2. Worksheets: problems from the book and R programming problems

- Weekly: ~10 total. **Important these are turned in on time!**
- You may discuss questions with other but
 - Work must be your own
 - List everyone you discussed things with

3. Final project: A chance to demonstrate that you can apply the central concepts in the class to analyze data

- ~10 pages, including code and description
- Class presentation on the project

Policies

Very important to turn your work in on time!

- Three strikes late policy
 - First late assignment you will get an email asking where your assignment is
 - Second late assignment you will be emailed a drop form
 - Third late assignment **you will not receive an evaluation for the class!**

Attend class

You can use a laptop for notes, but obviously do not use it for anything not related to the class

Check your Hampshire email and/or setup mail forwarding. I will send announcements to your Hampshire email account as well as Slack

Policies

Incompletes will only be given out in very exceptional circumstances

Special needs: please let me know if you have a disability or special need. Also it would be a good to talk to Aaron Ferguson – x5498

Academic dishonesty: you can work on the worksheets with others but the work you turn in needs to be your own (i.e., you need to understand the concepts).

Class survey and Slack

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

<https://goo.gl/GxE8qk>

Join the Slack group:

<https://https://goo.gl/DvMsNB>

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 and randomness

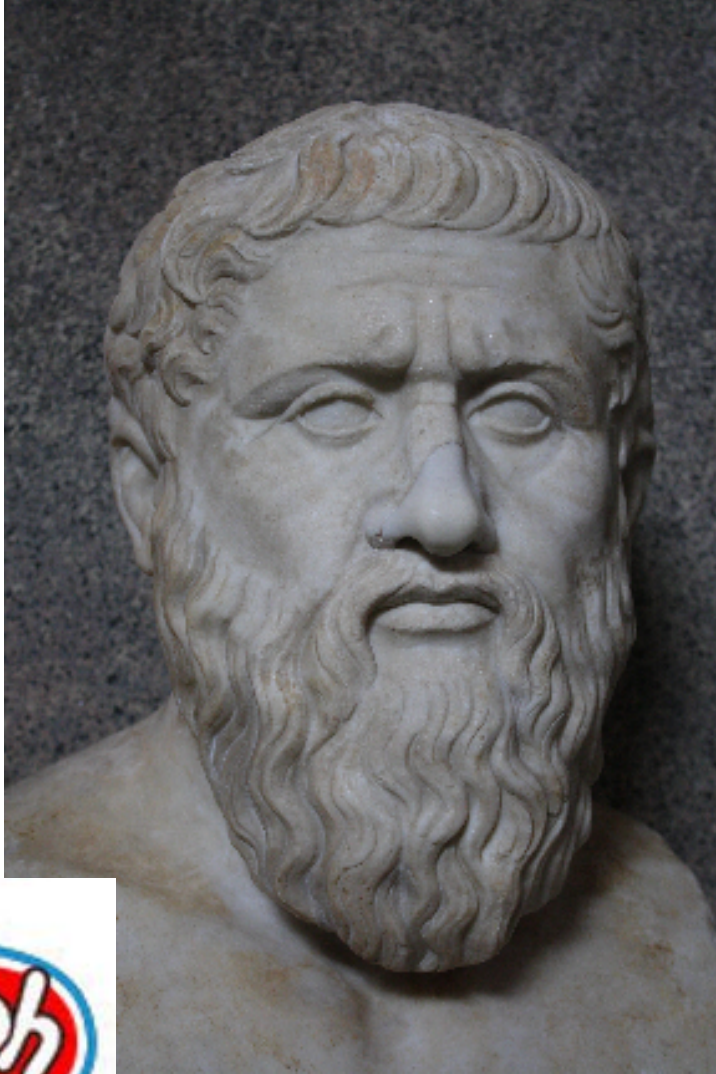
It's part of an argument

- 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



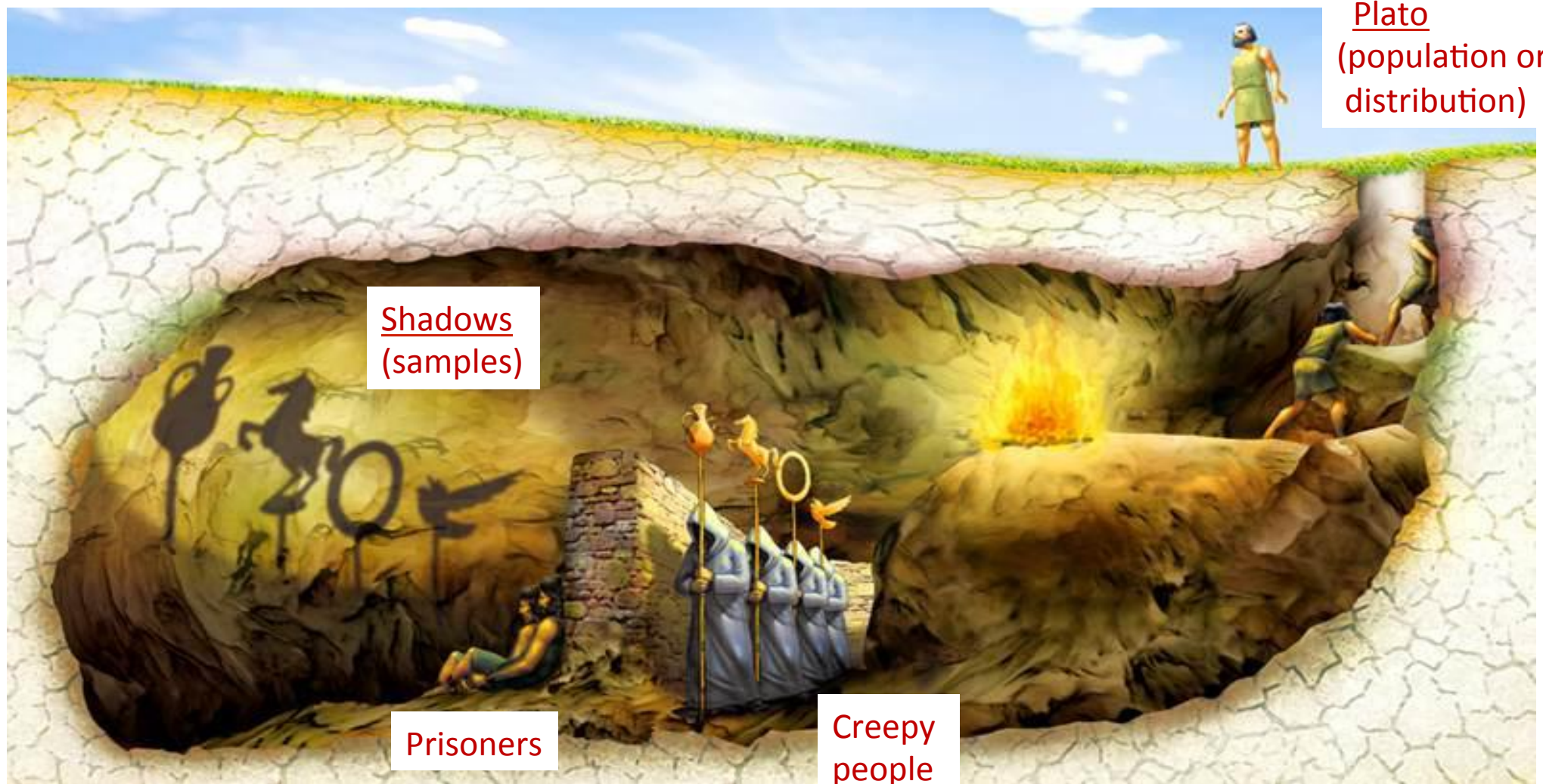
The truth!



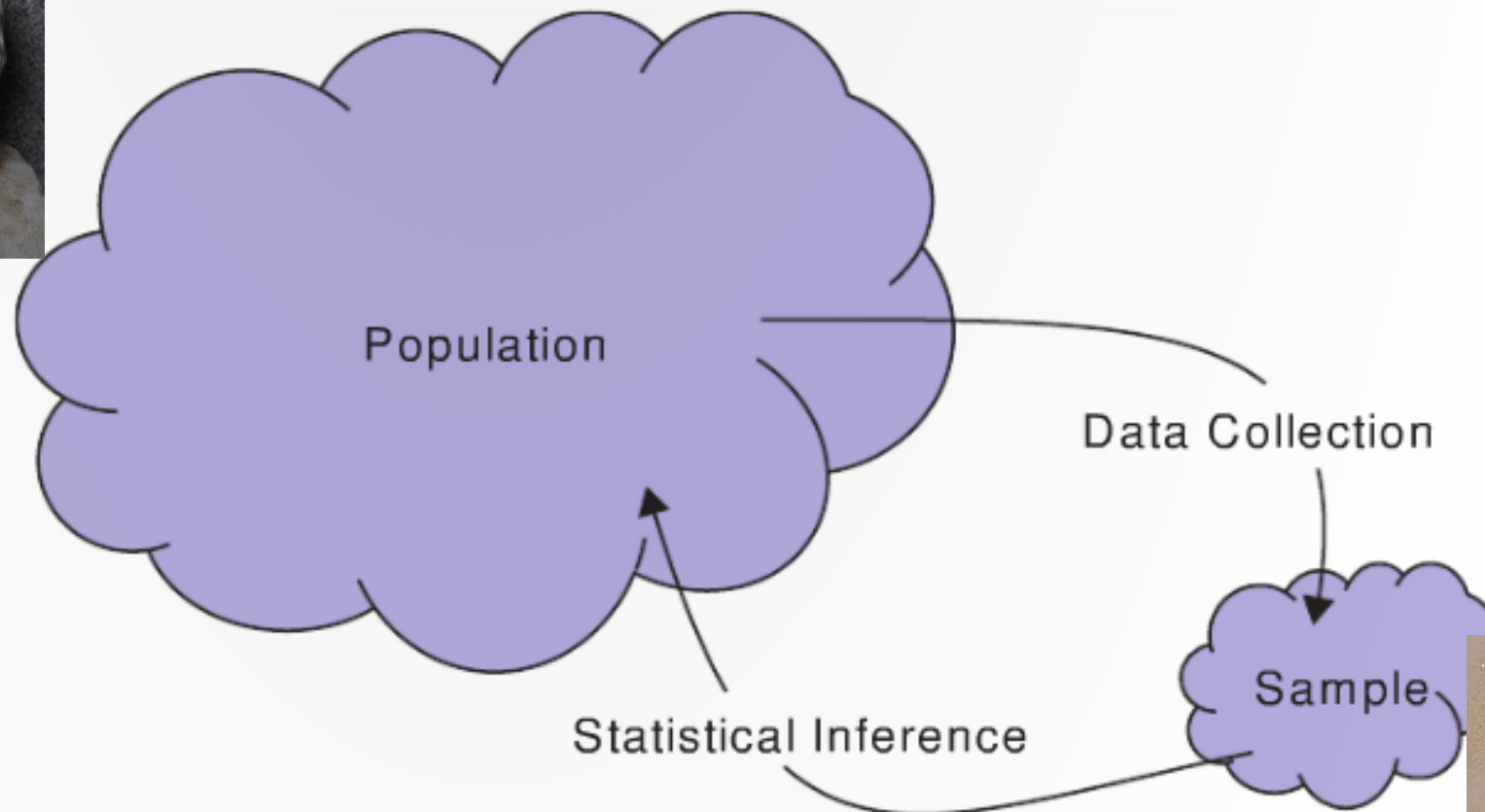
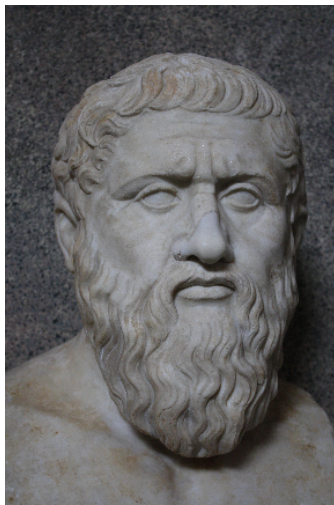
If we could see all the (infinite) data, we would know the truth!

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

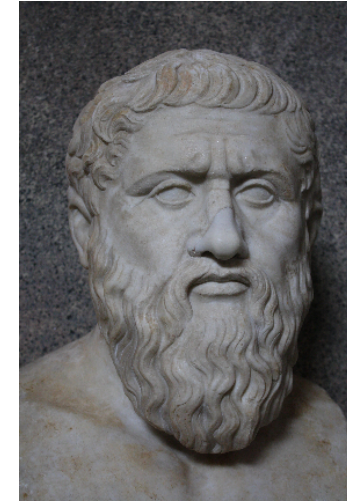


From The Republic (~ 380 BCE)



Sample from a Population

Population: all individuals/objects of interest



Sample: A subset of the population



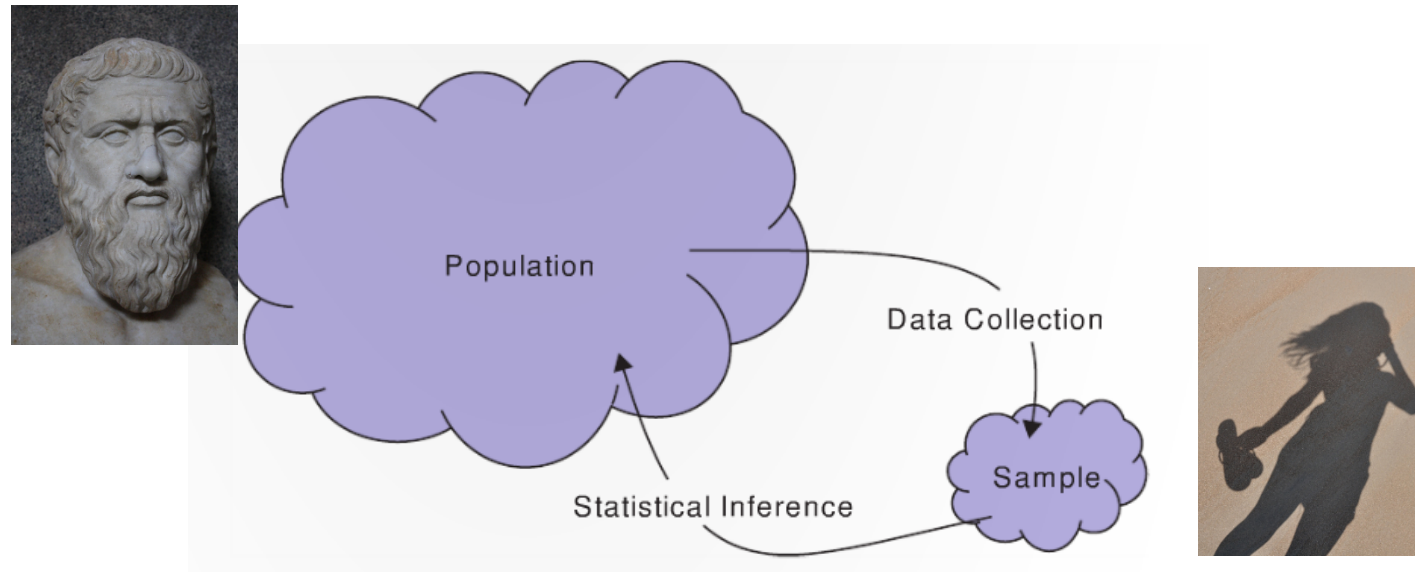
Descriptive and inferential statistics

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



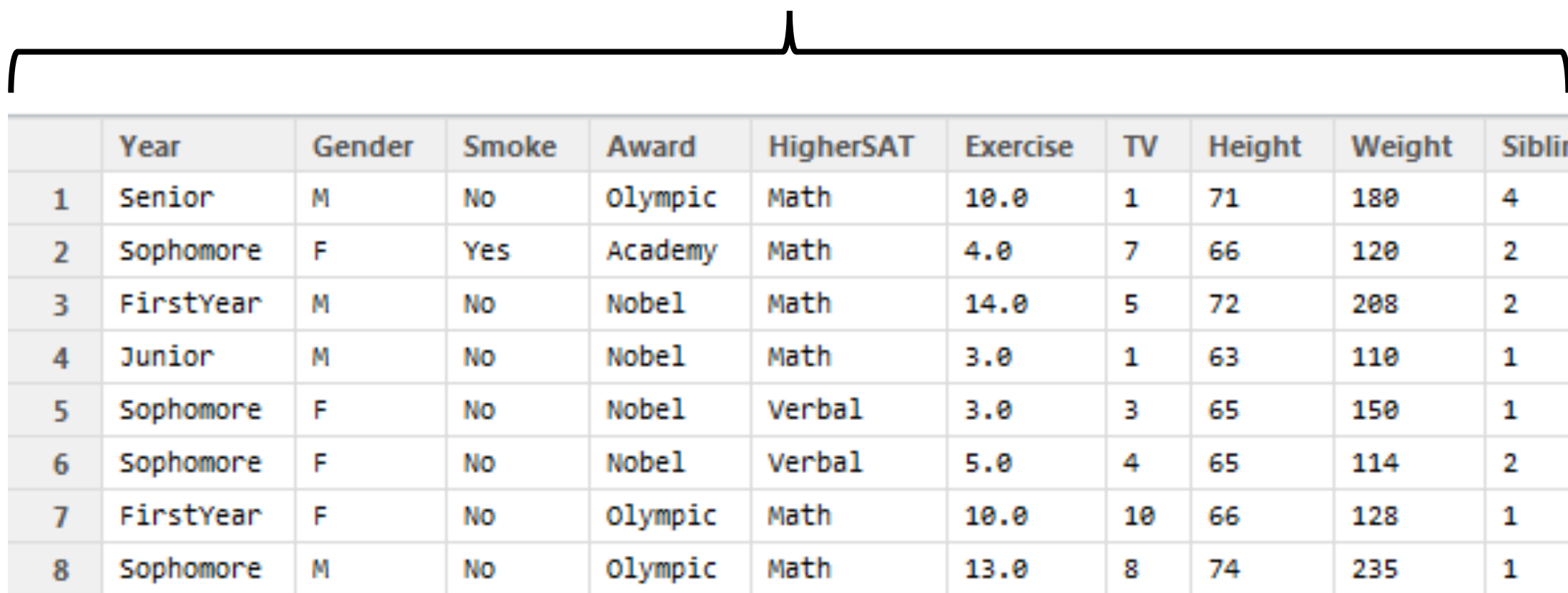
Structured data – exploring the shadows



An Example Dataset (Shadows)

Variables

Cases



	Year	Gender	Smoke	Award	HigherSAT	Exercise	TV	Height	Weight	Siblings
1	Senior	M	No	Olympic	Math	10.0	1	71	180	4
2	Sophomore	F	Yes	Academy	Math	4.0	7	66	120	2
3	FirstYear	M	No	Nobel	Math	14.0	5	72	208	2
4	Junior	M	No	Nobel	Math	3.0	1	63	110	1
5	Sophomore	F	No	Nobel	Verbal	3.0	3	65	150	1
6	Sophomore	F	No	Nobel	Verbal	5.0	4	65	114	2
7	FirstYear	F	No	Olympic	Math	10.0	10	66	128	1
8	Sophomore	M	No	Olympic	Math	13.0	8	74	235	1

An Example Dataset (Shadows)

Categorical Variable

Quantitative Variable

Cases
(observational units)

	Year	Gender	Smoke	Award	HigherSAT	Exercise	TV	Height	Weight	Siblings
1	Senior	M	No	Olympic	Math	10.0	1	71	180	4
2	Sophomore	F	Yes	Academy	Math	4.0	7	66	120	2
3	FirstYear	M	No	Nobel	Math	14.0	5	72	208	2
4	Junior	M	No	Nobel	Math	3.0	1	63	110	1
5	Sophomore	F	No	Nobel	Verbal	3.0	3	65	150	1
6	Sophomore	F	No	Nobel	Verbal	5.0	4	65	114	2
7	FirstYear	F	No	Olympic	Math	10.0	10	66	128	1
8	Sophomore	M	No	Olympic	Math	13.0	8	74	235	1

Edmunds transaction data

Discuss!

- 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

- <https://goo.gl/GxE8qk>

Practice problems from Lock 5, first edition:

1.1, 1.3, 1.5, 1.11, 1.25, 1.26

Chapter is posted on Moodle

You can check the answers at the end of the book

Question



Q: What programming language do pirates use?

A: Arrrr

Q: Worst joke of the semester?

A: Wait and see...

Log in to asterius

Accessing R with the Hampshire server:

<https://asterius.hampshire.edu>