Introduction

Grinnell College

Aug 23, 2025

My Background

In full: Reid Vincent Eli Paris, Ph.D.

- Melrose, IA, pop. 110
- University of Maine (BS)
- Iowa State University (MS, PhD)
- Three summers at ORNL
- Taught (and TA'd) for intro stat courses
- Stat consulting the last 5 years (140 projects)

My Research

Experimental Design

- How to get the biggest bang for your buck
- Dissertation on bias in experiments
  - Overly simple model due to resource limitations
  - Unrealistic experimental conditions
  - Hyundai and Nissian fined $230 million
- Order-of-Addition Experiments is my current interest
  - Deglaze your pan with....
    - ★ soy sauce then chocken stock *or*
    - ★ chicken stock and then soy sauce?
  - Ordering of medication matters

and random papers for consulting projects (concrete shrinkage, photoluminescence crystals, UV air filtration, etc...)

My Quirks

- Avid chess player dspite not being great
  - 62^{nd} percentile
  - Alekhine defence, always
- 21 National Parks and counting
  - Shout out to national forests and state parks though
  - Valley of Fire STATE Park all-time favorite
  - Summer desert camping
- DnD nerd
- Violin and mandolin
- Occasionally non sequitor statements but I dovetail back to a point
- Multiple audial sources is very distracting for me, please limit talking in lectures

Things We Will Study

A brief sketch of topics in the class

1. Describe data and variable relationships
   - graphical displays
   - designing studies
2. Estimation
   - Populations vs Samples
   - Confidence intervals
3. Hypothesis Testing
   - z-test
   - t-test
   - Chi-square tests (cut for time?)
4. Statistical Models
   - Regression

What are you learning today?

What is *statistics*, and why do we need it?

How would you describe the statistical framework to a relative?

What is an **observation** and how do we describe its characteristics?

What types of **variables** are there, and when is each appropriate?

What is Statistics?

Statistics is the study of variability

- Deterministic systems are what the mathematicians study
- How much difference (variability) is there between groups of things?

Eg Are girls the same height as guys?

- Variability due to age
- Variability due to environmental + genetic factors
- Variability due to biology
  - want to estimate this
- Variability due to random noise

Old Statistician Trick

Pick a number between 1 and 4

Two questions

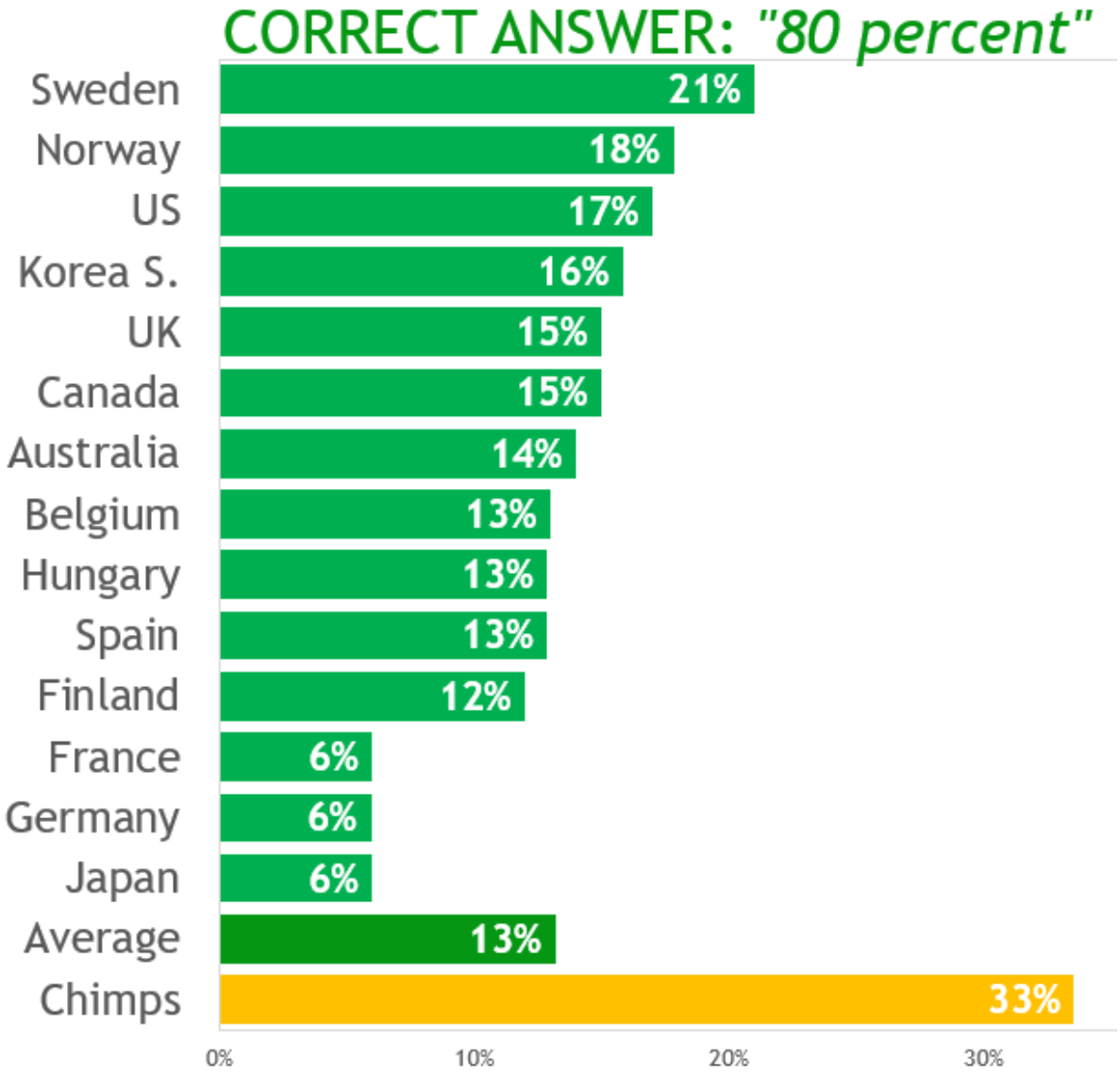
Question 1: What percentage of the world's 1-year-old children have been vaccinated against at least one disease?

A) 20%
B) 50%
C) 80%

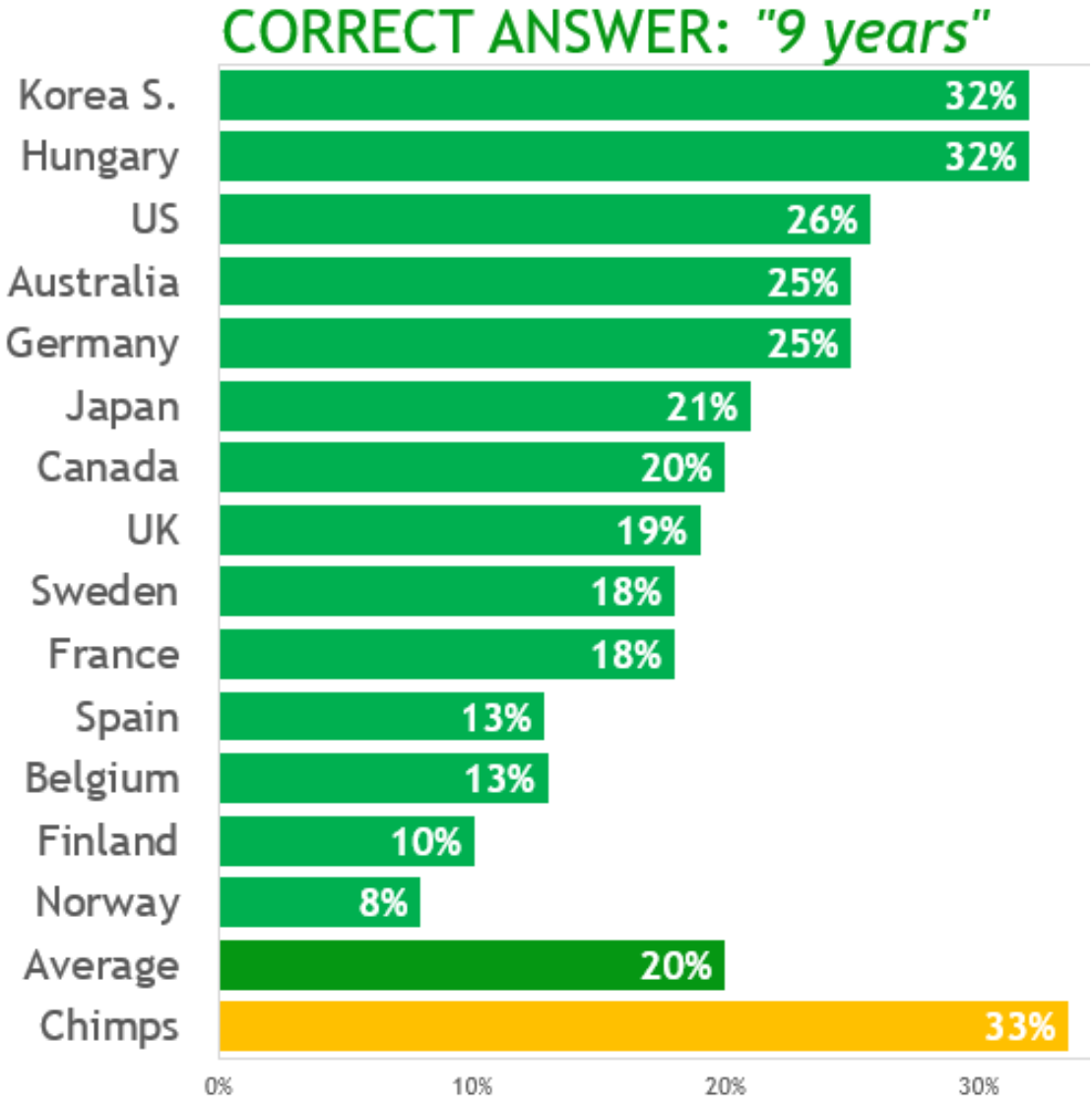
Question 2: Worldwide, 30-year-old men have 10 years of schooling, on average. How many years do women of the same age have?

A) 3 years
B) 6 years
C) 9 years

Vaccination

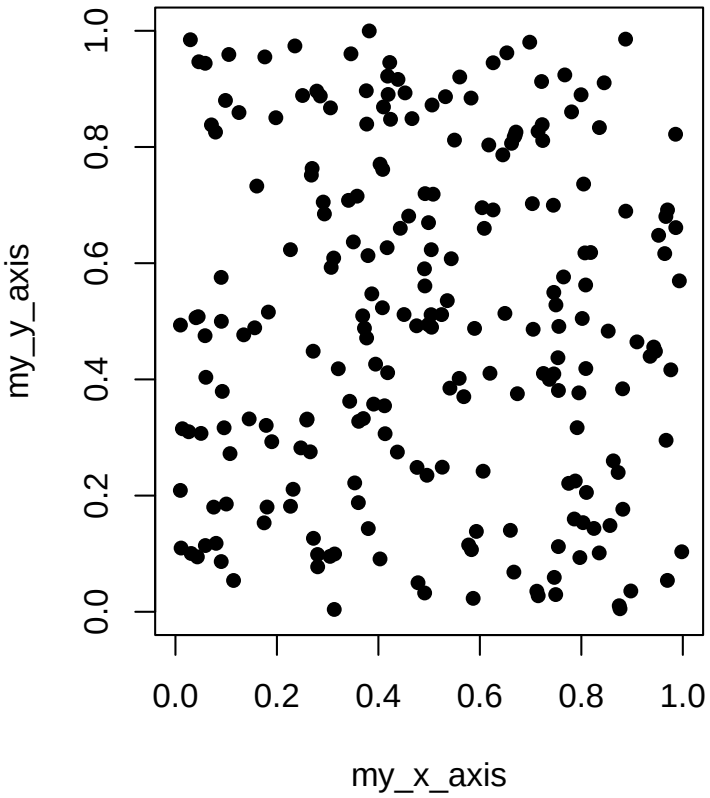


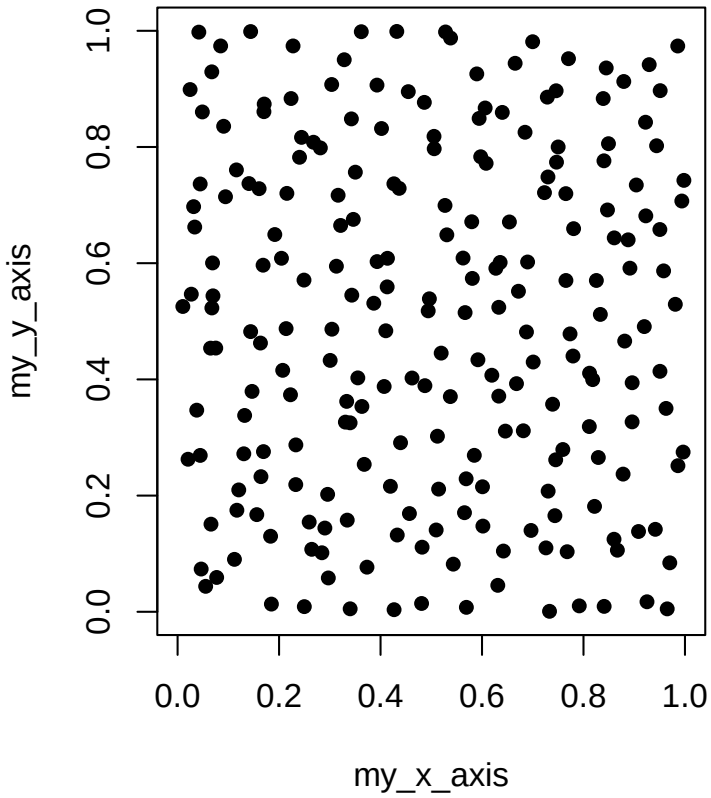
School


CORRECT ANSWER: "9 years"
Korea S.
32%
Hungary
32%
US
26%
Australia
25%
Germany
25%
Japan
21%
Canada
20%
UK
19%
Sweden
18%
France
18%
Spain
13%
Belgium
13%
Finland
10%
Norway
8%
Average
20%
Chimps
33%
0%
10%
20%
30%

Dots

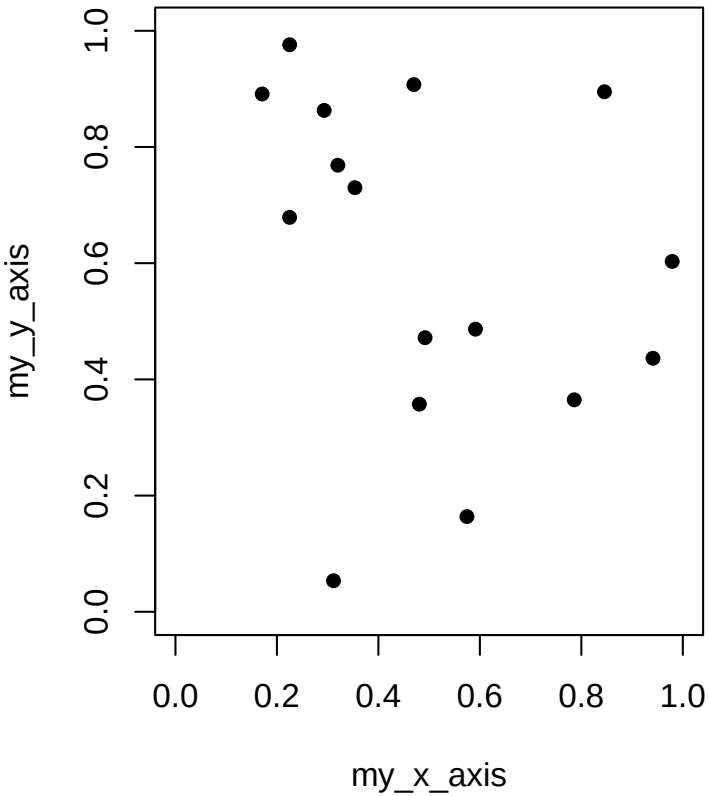
Which of these graphs look random?

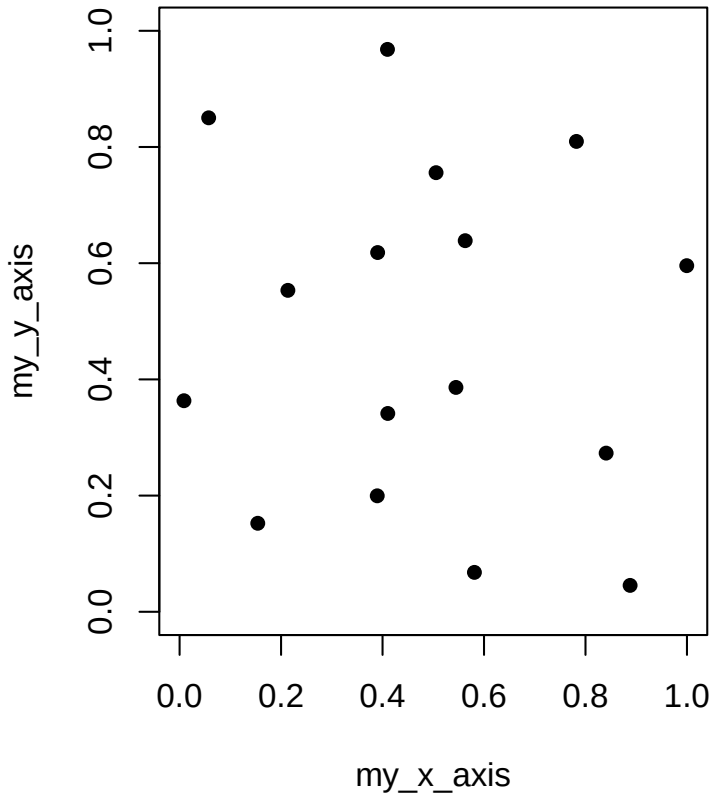




Dots 2

Sometimes it's easier if we simplify...





Why do we need statistics?

Human beings are great at identifying patterns

- Cognitive biases
- Poor intuition of uncertainty and randomness

Statistics gives us a framework for answering questions about the world using data (scientific method)

1. Clearly define your research question
2. Construct a hypothesis
3. Collect data
4. Consider evidence
5. Draw conclusions

Populations and Parameters

A **population** is a constrained group of subjects/events/things about which we wish to ask a scientific question

A **parameter** is a *quantifiable* attribute of a population. It is often assumed to be a fixed value within the bounds of the population

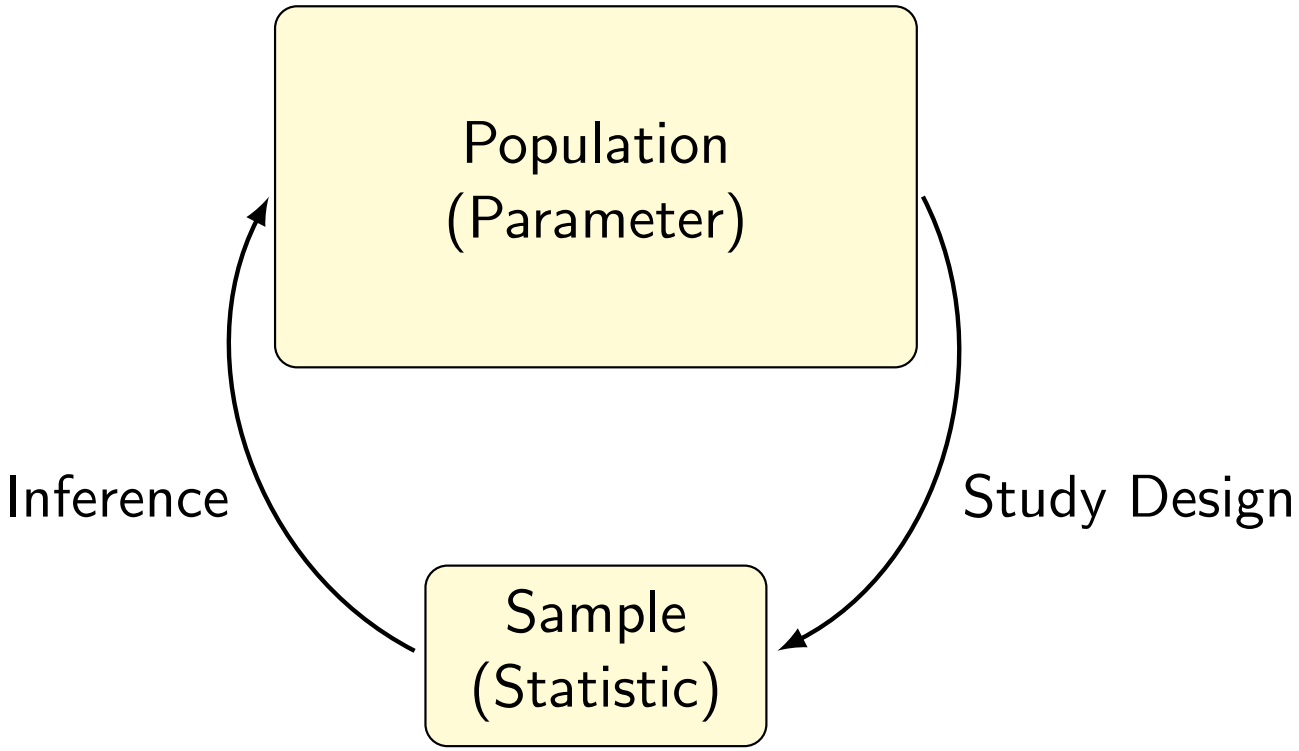
A **census** is a complete collection of data for a population. This lets us exactly determine the value of a parameter within the population

Samples and Statistics

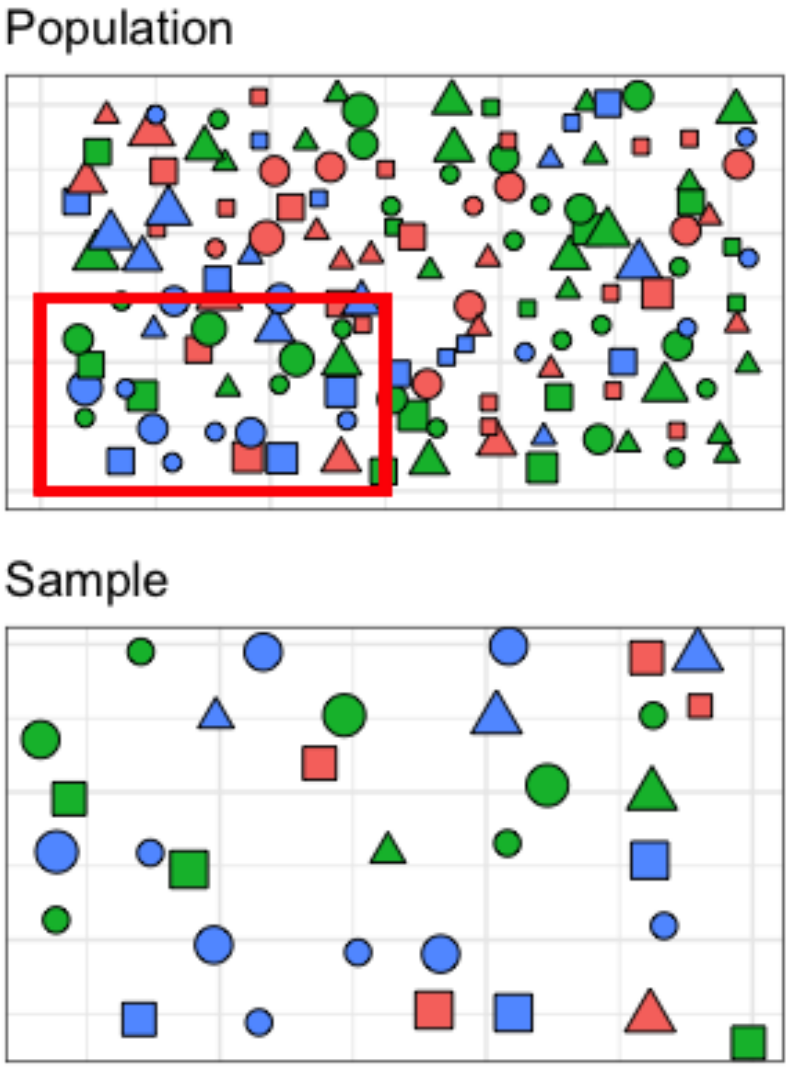
A **sample** is (often) a much smaller, (generally) *randomly collected* subgroup of a larger population

A **statistic** is an *estimate* of a parameter that we get using data collected from the sample

The Statistical Framework



Population and Samples



An example

Suppose we are interested in determining the average height of students currently enrolled at Grinnell College

Does it matter *which* students we sample?

Does it matter *how many* students we sample?

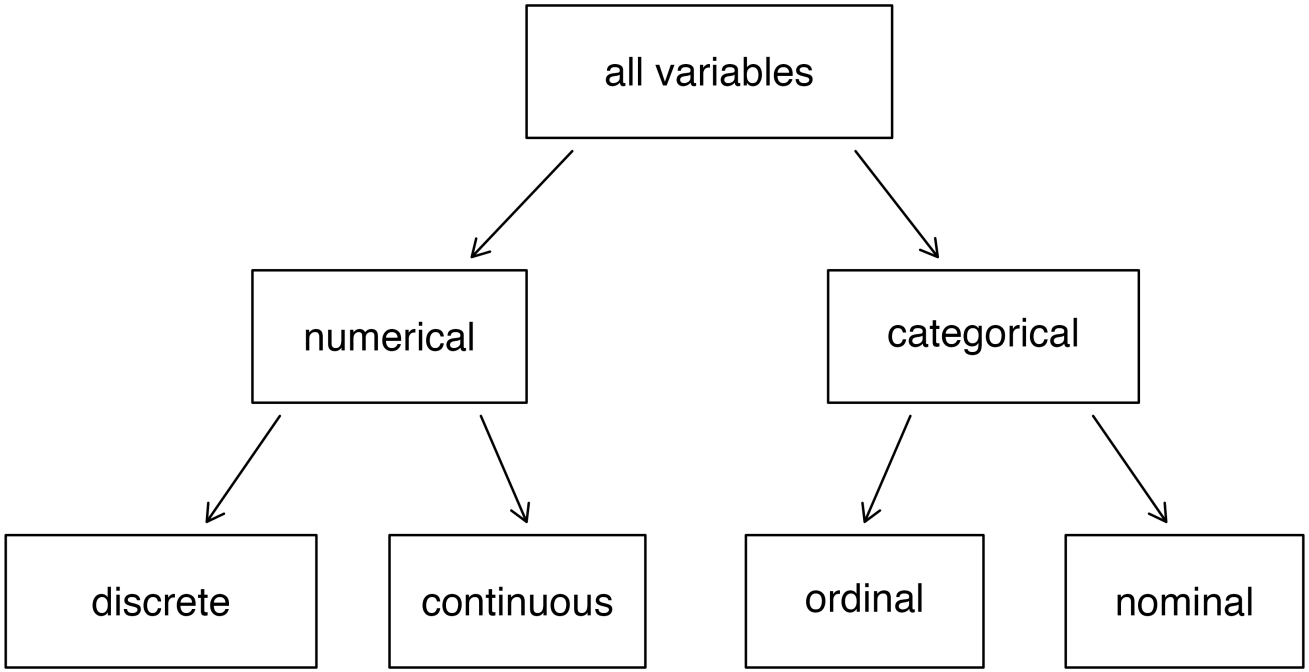
Some definitions

An **observation** (sometimes called an observational unit or case) is the literal subject/thing we are collecting data from

Characteristics of an observation are known as **variables**. Variables typically come in one of two types:

1. **Quantitative Variable:** Typically data that is stored in the form of *numbers*, and is numerical in nature
   - Continuous data i.e., height and weight
   - Discrete data (only specific values allowed) i.e., points scored in a game
2. **Categorical Variable:** variables that are naturally divided into *groups*
   - Binary (two groups)
   - Nominal (no ordering) ex: eye color
   - Ordinal (ordering makes sense) ex: year in college (F/J/So/Se)

Variables


all variables
numerical
categorical
discrete
continuous
ordinal
nominal

Example: Herbicde

A new computer system attempts to spray only weeds (see-and-spray) and not blanket the who field (full coverage)

Field ID	Spray Type	Wind Speed (mph)	Height (in)	Weeds covered (%)
A	see-and-spray	12	18	72
B	see-and-spray	15	18	55
C	full coverage	8	18	98
D	full coverage	11	18	97
E	see-and-spray	12	24	66
F	full coverage	14	24	95

Gray areas

The type of variable dictates how we analyze it:

- We often use the **mean** or **average** to analyze quantitative variables
- We often use **proportions** or **percentages** to analyze categorical variables

Sometimes there are situations in which a variable is technically one type, but it may be more useful to analyze it as another. Sometimes the type of variable can be different depending on how we record or organize our data.

Gray areas

Thoughts on if these might be used as quantitative or categorical variables?

1. Grades for a statistics class
2. A Likert Scale with five levels, measuring pain from "None at all" to "Extreme"
3. The year of birth for people enrolled in STA-209

"An approximate answer to the right problem is worth a good deal more than an exact answer to an approximate problem."
John Tukey, Statistician

Key Takeaways

- Populations are the things we want to study (students)
  - A parameter is a measurable summary of the population (mean height of students)
- A sample is a subset of the population (my students)
  - A statistic is a measurable summary of the sample (mean height my students)
- An observation is the smallest unit of study within a population (Jack's height)

Key Takeways 2

Variables primarily come in two types:

- Quantitative
  - Continuous (height)
  - Discrete (number of people)
- Categorical
  - Binary (disease status)
  - Nominal (favorite color)
  - Ordinal (educational attainment)

Closing

To Do Attempt to download R (first) and RStudio (second)

Suggested Readings Sections 1.2.1, 1.2.2, and 1.2.3 from IMS

Next Time: Introduction to R

Sources

IMS textbook

Prof.s Miller, Nolte and Friedrichsen's course notes

Dr. Ziegler's (ISU) course notes