# Introduction

Grinnell College

September 2, 2024

# Outline

A brief outline of 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
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 science and art of collecting and using data to learn about things

Statistics is about **variation**

- world is full of data
- these data exhibit variation (they aren't all the same)
- noticing, displaying, and quantifying this variation helps us learn
- end goal is to explain variation (why are things different?)

# Two questions

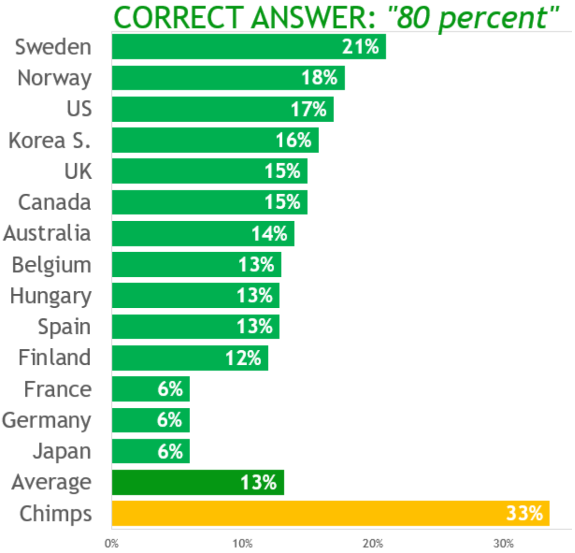
**Question 1:** What percentage of the world's 1-year-old children have been vaccinated against at least one disease?

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

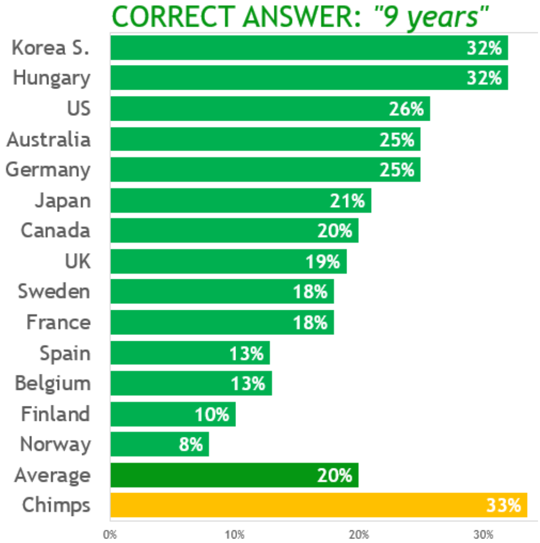
**Question 2:** Worldwide, 30-year-old men have 10 years of schooling, on average. How many years do women of the same age have?

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

# Vaccination

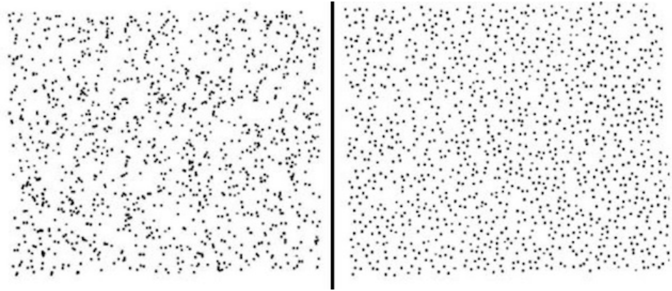


# School


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 boxes do you think reflects true randomness, and which of these seems artificially contrived?



# 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. Construct a hypothesis
2. Collect data
3. Consider evidence
4. 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

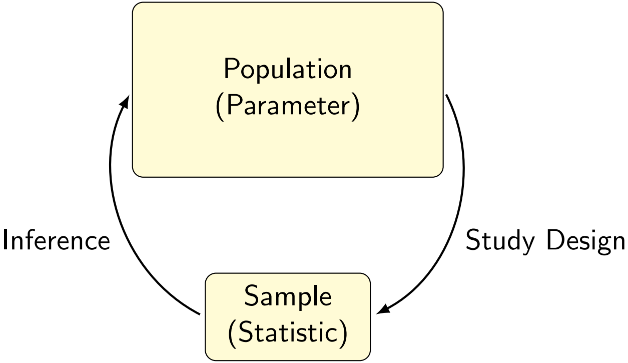
A **census** is a complete collection of data for a population. This lets us exactly determine the value of a parameter within the population

# Samples and Statistics

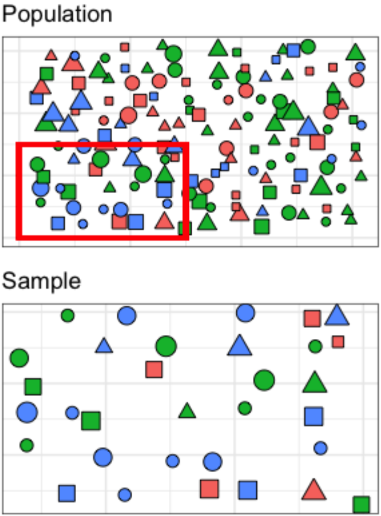
A **sample** is (often) a much smaller, (generally) *randomly collected* subgroup of a larger population

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?

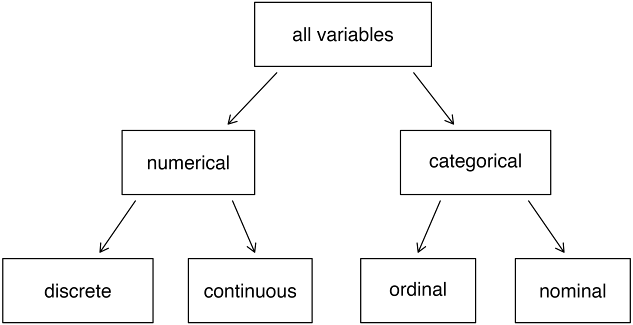
# Some definitions

An **observation** (sometimes called an observational unit or case) is the 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

# 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

Take a few minutes to discuss these questions with those around you, whether 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

- Statistics, as a discipline, gives us tools for analyzing variability in our data and answering scientific questions
- Parameters are quantifiable attributes of populations that we are interested in study. A sample is a subset of a population, and a statistic is an estimate of a parameter that we calculate using data from the sample
- An observation is the smallest unit of study within a population. It's characteristics are called variables
- Variables primarily come in two types:
  - Quantitative
    - Continuous (height)
    - Discrete (number of people)
  - Categorical
    - Binary (disease status)
    - Nominal (favorite color)
    - Ordinal (educational attainment)

# Knowledge Check

Why do we need statistics?

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?

# Summary

Statistics is a domain agnostic tool that allows us to make quantitative statements about a population

Most data that we encounter will be categorical or quantitative in nature

**Next Time:**

- Introduction to R
- Read Sections 1.2.1, 1.2.2, and 1.2.3 from IMS

# Sources

IMS textbook
Professor Miller's and Professor Nolte's course notes
Dr. Ziegler's (ISU) course notes