

# Statistics Introduction

Grinnell College

Fall 2025

# Outline

## A brief outline of the class

1. Describe data and variable relationships
  - ▶ graphical displays
  - ▶ designing studies
  - ▶ analyzing results of studies
2. Probability
  - ▶ Likelihood of things happening
  - ▶ Probability/Odds/Risk
  - ▶ Lots of fun gambling math
3. Estimating Things
  - ▶ Populations vs Samples
  - ▶ Confidence intervals
4. Hypothesis Testing
  - ▶ How to answer questions using data
5. Statistical Models
  - ▶ Describing relationships between things using data and math
  - ▶ Predicting values using information that we know

# What are you learning today?

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

How would you describe the statistical framework to someone else?

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?

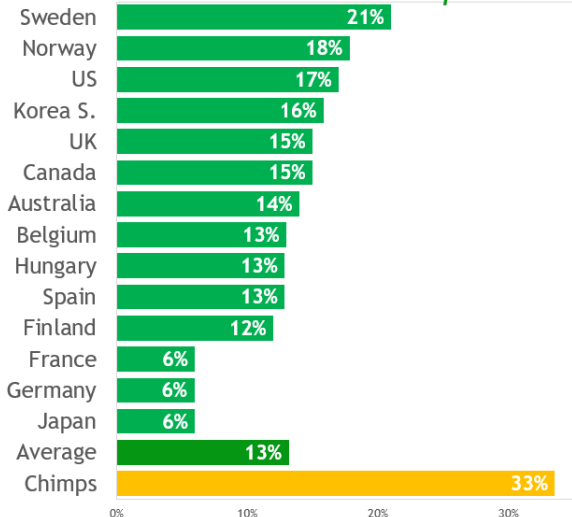
- A) 20%
- B) 50%
- C) 80%

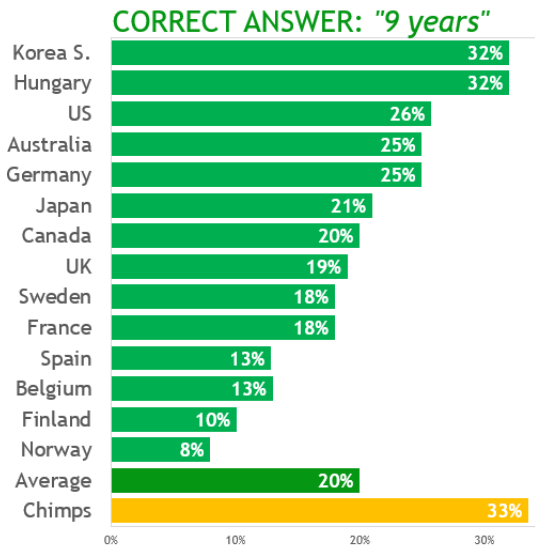
**Question 2:** Worldwide, 30-year-old men have 10 years of schooling, on average. How many years do women of the same age have?

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

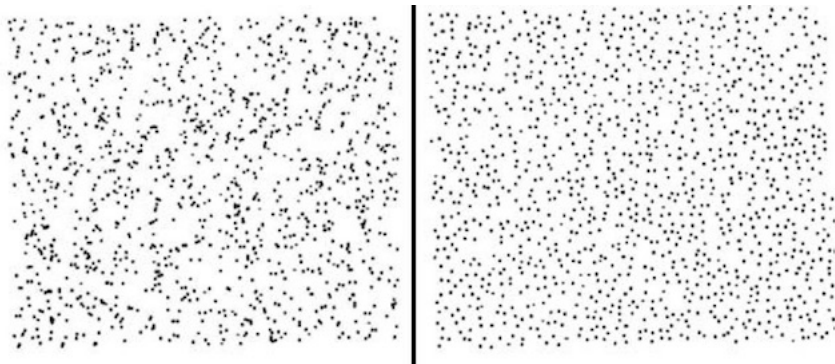
# Vaccination

**CORRECT ANSWER: "80 percent"**





# Randomness





# 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 (some statement we want to test)
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

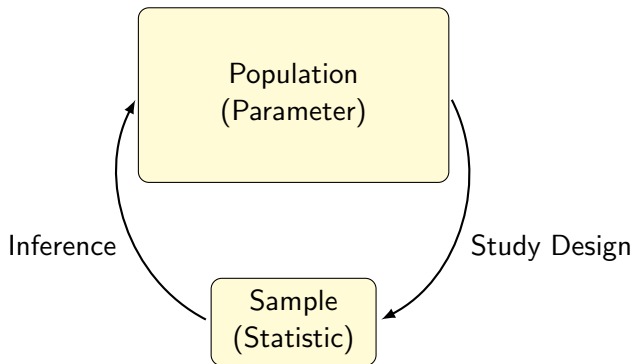
A **census** is a complete collection of data for a population. This lets us exactly determine the value of a parameter within the population

# Samples and Statistics

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

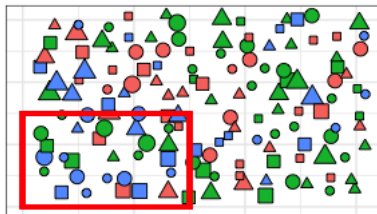
A **statistic** is an *estimate* of a parameter that we get using data collected from the sample

# The Statistical Framework

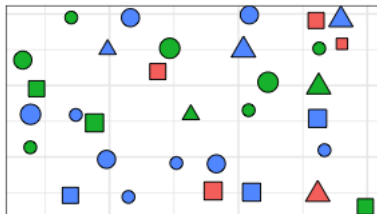


# Population and Samples

Population



Sample



# Populations and Samples – Example

Suppose we want to determine the average income of people in Iowa

Does it matter *which* people we sample?

Does it matter *how many* people we sample?

# Population and Samples – Soup

Suppose I want to know how good a big pot of soup tastes. Do I need to go and eat the entire soup to answer this? No!



As long as the soup is mixed well, one bowl of soup will give us a good idea of how the entire thing tastes.

- Population = entire bowl of soup
- Sample = one bowl of soup
- If something is missing, we won't learn about it (underdone potatoes)

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

- these are the individual things in our sample

Characteristics of an observation are known as **variables**.

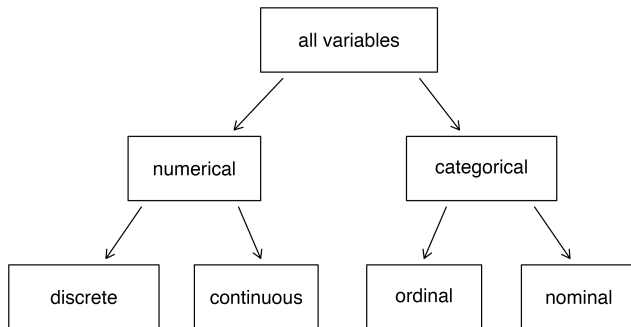


# Variables

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 (limited by precision)
  - ▶ Discrete data (only specific values allowed) i.e., points scored in a game
2. **Categorical Variable:** variables that are naturally divided into *groups*
  - ▶ Ordinal (ordering makes sense) ex: grade year in school
  - ▶ Nominal (no ordering) ex: eye color

# Variables



# Variables – Example

1. **Marvel Cinematic Universe films.** The data frame below contains information on Marvel Cinematic Universe films through the Infinity saga (a movie storyline spanning from Ironman in 2008 to Endgame in 2019). Box office totals are given in millions of US Dollars. How many observations and how many variables does this data frame have?<sup>8</sup>

	Title	Length		Release Date	Opening Wknd US	Gross	
		Hrs	Mins			US	World
1	Iron Man	2	6	5/2/2008	98.62	319.03	585.8
2	The Incredible Hulk	1	52	6/12/2008	55.41	134.81	264.77
3	Iron Man 2	2	4	5/7/2010	128.12	312.43	623.93
4	Thor	1	55	5/6/2011	65.72	181.03	449.33
5	Captain America: The First Avenger	2	4	7/22/2011	65.06	176.65	370.57
...	...	...	...	...	...	...	...
23	Spiderman: Far from Home	2	9	7/2/2019	92.58	390.53	1131.93

Data frame = arrangement of data to more easily read values

- Information for any 'case' is all in one row
- Variables correspond to the columns (Hrs, Mins, Release Date, etc.)

Why spend time on classifying variables?

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 age of people in this room

*"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
    - ★ Nominal (favorite color)
    - ★ Ordinal (educational attainment)

# Knowledge Check

Why do we need statistics?

How would you describe the purpose/goals of statistics to someone else?

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 or describe relationships between things

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

## **Next Time:**

- Making displays of data to visualize information
- Read Sections 1.2.1, 1.2.2, and 1.2.3 from IMS