

# Introduction to Statistics

Grinnell College

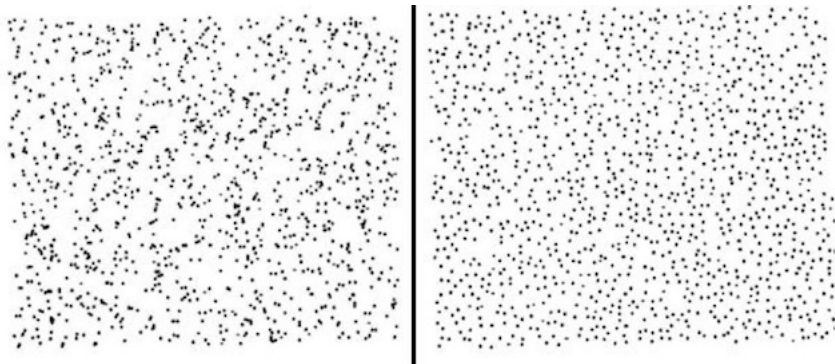
August 29, 2025

## A brief outline of the class

1. Part 1 – Data Summaries
  - ▶ Visualizations
  - ▶ Numerical Summaries
  - ▶ Tables
2. Part 2 – Basics of Hypothesis Testing
  - ▶ Study design
  - ▶ Samples and distributions
  - ▶ Hypothesis testing
3. Part 3 – Statistical Tools and Applications
  - ▶ Multivariate tests
  - ▶ Statistical modeling

# Dots

What differentiates these two distributions of dots? Which of these 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 understanding of uncertainty

**Statistics** as a discipline is about the *quantification of uncertainty*.

1. Construct a hypothesis
2. Collect data
3. Consider evidence
4. Draw conclusions

# Populations and Parameters

A **population** is a constrained set of events or subjects 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 or immutable quality within the bounds set by the population

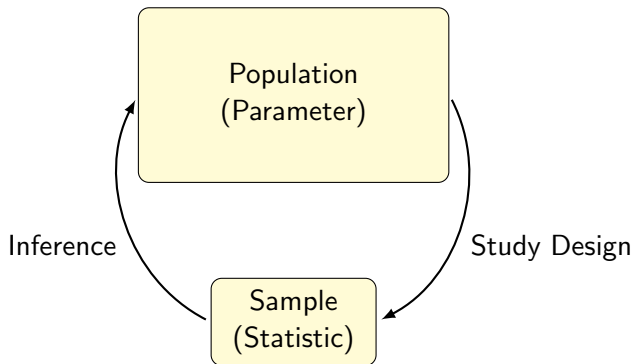
To determine the value of a parameter within a population with certainty is to conduct a **census**

# Samples and Statistics

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

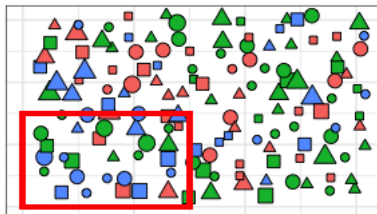
A **statistic** is an *estimate* of a parameter derived from data collected within the sample

# The Statistical Framework

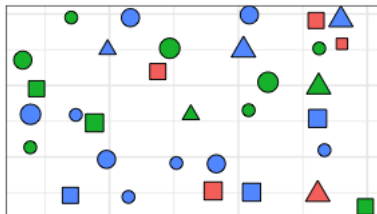


# Population and Samples

Population



Sample





## 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?

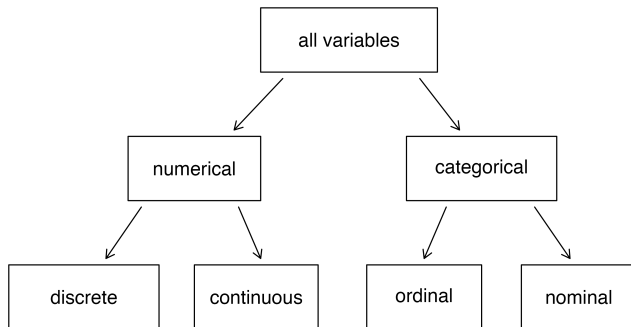
How much confidence do we have that our estimate of the average height is close to correct?

# Some definitions

In this course we will primarily be working with data derived from **observations**, our most basic unit of study. Characteristics of an observation are known as **variables**. Variables typically come in one of two types:

1. **Quantitative Variable:** Data that is typically stored in the form of *numbers* and is numerical in nature
  - ▶ Continuous data i.e., height and weight
  - ▶ Discrete data i.e., points scored in a game
2. **Categorical Variable:** Variables that are naturally divided into *groups*
  - ▶ Binary
  - ▶ Nominal
  - ▶ Ordinal

# Variables



# 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

# Key Takeaways

- Statistics is a domain agnostic tool that allows us to make quantitative statements about a population based on the properties of a sample
- Parameters are attributes of populations that we are interested in study. A sample is a subset of a population, and a statistic is a derived estimate of a parameter
- An observation is the smallest unit of study within a population. It's characteristics are called variables

# Key Takeaways

- Variables primarily come in two types:
  - ▶ Quantitative
    - ★ Continuous (height)
    - ★ Discrete (number of people)
  - ▶ Categorical
    - ★ Binary (disease status)
    - ★ Nominal (favorite color)
    - ★ Ordinal (educational attainment)