Introducing Data

1 Why Study Statistics?

- Many students taking STAT 103 usually are just trying to fill a course requirement. But this class can be a lot more than that. Statistics is not just a field for math and science majors. It can give you skills that can be applied to nearly any area of your life, regardless of what you're studying or what career you want.
- At some level, we are all interested in understanding the world around us. We naturally build our own mental models of reality; by default we are engaging in the same process as philosophers or scientists. We are constantly trying to establish what's true and what's not, what we should believe and what we should not believe.
- For example:
  - Is climate change real?
  - Does social inequality exist at a systemic level?
  - Is one type of treatment for cancer more effective than another?
  - What percentage of students with my major get the job that I'm looking for after graduation?
  - What kind of salary can I expect after college?
- Statistics is a system that we can use to reliably answer or make judgments about the questions we about the most, questions about ourselves, questions about the world and our place in it.
- This course has been built around the idea of relevance to our lives. We will repeatedly be connecting the topics we learn to things most of tend to care about. STAT 103 more of a critical thinking course dressed up in Statistics.

2 Types of Variables

- Statistics is the study of data. The first step in understanding data is to understand the different types of data you will encounter.
- Almost always, our data sets will consist of characteristics of people or things (such as gender and weight). These characteristics are called **variables** because they have variability, meaning they can change from sample to sample or person to person.
  - Something to note : In statistics, variables are not "unknowns" that you solve for like those you studied in algebra.

Two Types of Variables

- **Numerical variables (quantitative)** describe a quantity or a measurement. The values will be numbers.
- **Categorical variables (qualitative)** describe a quality. These values will be categories.

Numerical Variable	Categorical Variable
• Your weight, height, or age. • The temperature outside. • The distance or length of something.	• Your gender, hair color, or birth month. • Your favorite color. • If you own a pet or not.

- Identify the following variables as either numerical variables or categorical variables. We will denote categorical variables as C and numerical variables as N for the following few examples.
  - Your eye color C
  - The inches of rainfall during the month of April N
  - The number of people who like vanilla ice cream versus chocolate ice cream N
  - Your zip code C
  - Your yearly salary N
  - The time it takes you to travel to work N
  - Your ethnicity C
  - The number of people who are Hispanic versus Asian N
  - The total ounces of coffee a person drinks in a day N
  - The number of miles driven on a car N
- Typically (but not always) variables that have a **unit of measurement** are numerical. Example: Inches, dollars, percent, feet, degrees Fahrenheit, ounces, miles, hours, minutes, etc.
- Categorical variables do not have a unit of measurement. They typically rely on opinion (do you like something or not), possession (do you have something or not), yes or no responses, etc.

Discrete and Continuous Variables

We can further categorize numerical variables! There are two types of numerical variables.

- **Discrete variables** : Numerical values that you can list or count.
  - Example : The number of people in your class. The number of siblings you have. The number of cars in your driveway.
- **Continuous variables** Numerical values that occur over a range.
  - Example : Your age. Scores during a season. The number of steps on your fitbit throughout a day.

To help us better make the distinction between discrete and continuous variables, we can think of discrete variables being some value, usually a whole number, that won't change in the specified range or time. On the other hand, discrete variables will change over a specified range or time.

Example : Consider the variables below. Identify them as continuous or discrete. We will denote continuous as C and discrete as D for the following few examples.

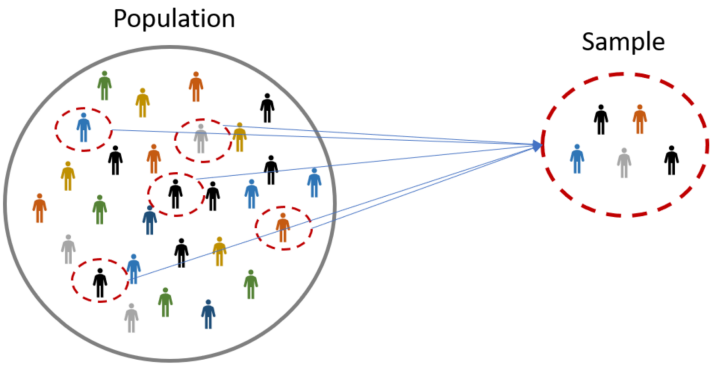
- The time elapsed from when you left your house to when you arrived on campus. C
- The number of people who got in a car accident in 2021. D
- Blood alcohol level of a an individual who has been drinking tonight. C
- The height of a person each year, starting from infancy. C
- The amount of avocados an avocado tress produces in one year. D

3 Conducting Studies

Samples and Populations

Now that we have identified the different types of variables we will explore, we can explore the beginning stages of how a study is conducted. Studies usually require us to observe something about a population of interest. But since we can't always observe an entire population, we tend to observe samples from the population.

- If you look at the picture below, you will notice that a **sample** is a small part of the entire **population.**
- Think of a population as containing **everything** you are studying whereas the sample is a small part of that large population. You can take as many random samples as you'd like from one population.



Example 1: A researcher wants to estimate the average height of all women aged 20 years or older. Hence the researcher selects 45 women at random and obtains an average height of 63.9 inches. What is the population and what is the sample?

Population is all women aged 20 years or older while sample is the 45 women that are selected for the study.

Example 2: Researchers claim that 70% of all children, 10 years and younger, consume too much sodium. To study this phenomenon, a nutritionist randomly selects 75 children 10 and under. What is the population and what is the sample?

Population is all children, 10 years and younger whereas sample is 75 children who were selected who were 10 and under.