Chapter 1: The Language of Statistics

1 Why study statistics?

- Analyzing statistics is like a microscope that allows us to see things that are too big for our eyes to take in. It allows us to take a big thing and shrink it down to something small so that we can see it. We'll never really be able to comprehend the vast complexities of a really large set of data. Trying to gain that kind of perspective is immediately overwhelming. Statistics gives us tools to analyze tendencies, trends, likelihoods, and associations.

  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.

2 Observational units and variables

Do you think there is a relationship between a student's phone use in class and their grades?

- Yes
- No

Are there any details about "phone use" that are important to consider?

Yes, phone use may have a negative effect on your grades because it makes you more distracted, but how you're using your phone might matter. After all, you could be using it to look up information for class or contact other students to schedule study time or ask for help!

- A dataset contains information about a group of individuals/things or observational units. The characteristics of these observational units are recorded as variables.

  For example, the researcher collecting data on student phone use might ask individual students to report the number of times they checked their messages during class. In this case variable is the number of times messages were checked during class and observational unit is one student

- A variable is classified as qualitative or categorical if it places an individual into one of several groups; it is classified as quantitative if it takes numerical values that can be used in arithmetic.

- There are **two types of quantitative variables.** A discrete variable takes a fixed set of possible values, and it is not possible to get any value in between. In contrast, the range of outcomes for a continuous variable includes an infinite number of possible values.

- 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.

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 Statistical question

A statistical question will ask a question in which the answer can change; data must be collected and analyzed, and the answer offers an explanation of the information. In other words, a good statistical question should anticipate variability in the data and do not have an exact answer. A good statistical question seeks to find a general trend, central tendency, or typical value.

Example: You want to analyze the heights of students in your introductory statistics class using statitsics. Which of the following is a good statistical question?

- How tall are you?
- How tall, in inches, was Devon on his alst birthday?
- Is Mia taller than 60 inches?
- How tall are students in your class?

The correct answer is How tall are students in your class? as this question anticipates variability in the data and do not have an exact answer.

Example 2: You want to study the effect of different teaching methods on student performance in a statistics course. Which of the following is a good statistical question?

- How effective is the lecture method for teaching statistics?
- How many students preferred the interactive teaching method last year?
- Do students perform better with the lecture method or the interactive method?
- What are the performance differences between students taught with the lecture method and those taught with the interactive method?

The correct answer is What are the performance differences between students taught with the lecture method and those taught with the interactive method? as this question anticipates variability in the data and does not have an exact answer.

Try it yourself: You want to analyze the impact of socioeconomic status on students' access to online learning resources. Formulate a good statistical question to survey the aforementioned problem.

