Chapter 1

Yu Yang

School of Statistics
University of Minnesota

September 9, 2022

What is Statistics?

Statistics is the art and science of collecting, organizing, interpreting, and learning from data.

Three Aspects of Statistics

1. Design: Planning how to obtain data to answer the question of interest.
2. Description: Summarizing the data that are obtained.
3. Inference: Using sample data to learn about the population and to answer the statistical question.
   - 3.1 Parameter estimation
   - 3.2 Hypothesis testing
   - 3.3 Modeling

Course goal

Learn how to use statistical methods to translate data into knowledge so that we can investigate questions in an objective manner.

Examples:

1. (Parameter estimation) How can we estimate the average age of all students at the U?
2. (Hypothesis testing) Is there any significant difference between the salaries in two companies?
3. (Modeling) What is the relationship between the amount of time spent studying and the score received on an exam?
4. ...

Definition

- Population: the *population* is a collection of units of interest.
- Subject: *subjects* are the individual units of a population.
- Sample: a *sample* is the subset of the population for whom we have (or plan to have) data, often randomly selected.
  NOTE: Very rarely can we observe the *entire* population of interest. The basic goal of statistics is to:

  instead, observe a sample and use it to learn about the population.

Definition

- Parameter: a *parameter* is a number that describes a *population*. It is usually UNKNOWN.
- Statistic: a *statistic* is a number that describes a *sample*. It can be computed from data; therefore, it is KNOWN once a sample is obtained.

Estimate the average age of our class

- Population: a class of STAT 3011.
- Subject: each student in that class.
- Sample: those who filled in the google form
- Parameter: the true avearge age μ of our class. It is UNKNOWN.
- Statistic: $\overline{Age} = \frac{Age_1 + Age_2 + \cdots + Age_{94}}{94}$. It is KNOWN

Estimate the average age of our class

- Population: a class of STAT 3011.
- Subject: each student in that class.
- Sample: those who filled in the google form
- Parameter: the true avearge age μ of our class. It is UNKNOWN.
- Statistic: $\overline{Age} = \frac{Age_1 + Age_2 + \cdots + Age_{94}}{94}$. It is KNOWN

Question

Can we use this dataset to estimate the average age of all U students?
-**No, because STAT 3011 is intended for undergraduate students.**

What makes a "good" sample?
***It should be representative of the population*. This can be obtained by selecting sample subjects *randomly* (more in Chapter 4).**

More example

We want to know the average height of all students at the U. It is logistically impossible to measure everybody. Instead, we take this class as a sample, measure our heights, and average them.

- population:
  all U students
- sample:
  this class
- parameter:
  average height of all U students
- statistic:
  average height of this class

More example

Suppose we want to know what percentage of Minnesota adults own a firearm. Since it's impossible to ask all adult Minnesotans, we instead take a poll of 1000 Minnesotans by selecting a sample from the phone book.

- What is the population?
  All adult Minnesotans.
- What is the sample?
  The 1000 Minnesotans selected from the phone book.
- Is this a good sample?
  No. Some people don't have phones, have unlisted phone numbers, or have cell phones.