STAT 107 – Spring 2021

[Unit 1] Statistics & Statistic

Ha Khanh Nguyen (hknguyen)

What is Data Science?

"Data science is an inter-disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from many structural and unstructured data." (Wikipedia)

In simple terms, data science is the science of collecting and analyzing data fast and efficiently.

That's why in STAT 107, we learn both Statistics and Programming (Python)!

What is (the study of) Statistics?

What is a statistic?

Types of Statistical Studies

- Descriptive Statistics
  - Researchers aren't concerned with making generalizations to the greater population.
  - Have data from all or much of the population. We feel confident making statements about the whole population using our data (think a census).
- Inferential Statistics
  - When we want to understand something about a particular population, but we're unable to collect data from a significant portion of the population.
  - We then collect data from a sample of this population that we believe is *representative* of the population and can use our sample data to make an inference.
  - An **inference** is a conclusion we draw about the population based on information we have gathered from our sample.

Statistical Research Questions

A Research Question is a specific, directed question that we want to use data to help us answer. In our class, we are interested in answering research questions with data and statistics.

- Descriptive Questions:
  - "What temperatures does Champaign typically see during the year?"
  - "What percent of residents are in favor of Proposition 5?
- Associative Questions:
  - "On average, is the mortality rate in Champaign higher on days with higher temperatures?"
  - "Are homeowners more likely to support Proposition 5 than other residents?"
- Causal Questions:
  - "Does taking this medication cause increased risk heat stroke?"
  - "Does watching this ad affect the likelihood of a resident voting for Proposition 5?"
- Existence vs. Magnitude Questions:
  Asking "Is there an effect/relationship/difference...?" is not the same as asking "How much effect/relationship/difference...?" One asks about the existence of an effect, while the other asks about the magnitude of that effect.

Basic Terminologies in Statistics

Dataset vs. Variable

- A **dataset** is often in the form of a table, where each row represents one person/unit/observation, and each column represents one question/characteristic/type of measurement taken.
- A **variable** is a characteristic of interest we gather from each participant/unit through a question or measurement (one specific column of your dataset).

Population vs. Sample

- A **population** is the entire group we have an interest in learning and making an inference about.
- A **sample** is a subset of a population. We often use *n* to denote the sample size.

Parameter vs. Statistic

- A **parameter** is a numerical value that describes some characteristic about the population.
- A **statistic** is a numerical value that describes some characteristic about a sample.

Practice: Gallup conducted a poll to gauge the opinions of Adult U.S. Residents about gun laws. Gallup contacted a representative sample of 1,526 people. Among several questions asked, one asked about whether or not you supported a complete ban on individual gun ownership. 29% said yes.

Our population is..._____

Our sample is..._____ n =_____

Our variable of interest is..._____

The sample statistic we gathered is..._____

Do we know what the population parameter is? _____

To-do:

- Install Python using the Mac OS X Guide or the Windows Guide.
- Set up git for discussion section on Wednesday/Thursday using the Setting Up git Guide.

This lecture notes were adapted from "Course Notes for STAT 100: Statistics" by Kelly Findley & Ha Khanh Nguyen.