Notes

Week 1

Module 1 Week 1A

Introduction

- **Data Analytics** - using data to inform decisions.
- **Statistics** - calculations derived from a dataset used to convey important features of the data in a concise way; the science focused on collecting, analyzing, interpreting, and presenting data.
- Statistics can be used on both accurate/non-accurate data, and approriate *and* inappropriate techniques can be used; see **Stamp's Law** as a reminder to be cautious and critical when interpreting data.
- **Stamp's Law** - "The government are very keen on amassing statistics. They collect them, add them, raise them to the nth power, take the cube root and prepare wonderful diagrams. But you must never forget that every one of these figures comes in the first instance from the village watchman, who just puts down what he damn pleases." – Josiah Stamp, 1st Baron Stamp
  - **GIGO** - Garbage In - Garbage Out
- 2 Main Branches of Statistics:
  - **Descriptive Statistics** - describe the characteristics of a *particular dataset.*
    - e.g. average amount of debt held by 100 U.S. college students
  - **Inferential Statistics** - uses a *sample* of data to draw conclusions about and underlying *population* of interest.
    - e.g. using the sample of 100 U.S. college students to draw conclusions about all college students in the U.S.

    - Terminology
      - **Population** - collection of objects under study; defined by an analyst
        - **Parameter** - a numerical value describing a characteristic of a population.
      - **Sample** - the subset of the population we have data on.
        - *Probability Theory* - the mathematics of randomness, provides the link between a sample (what is observed) and a population (what we study, but is unobservable)
        - **Statistic** - a numerical value describing a characteristic of a sample.
      - In *Inferential Statistics*, we use a *Statistic* to draw conclusions about a population *Parameter* of interest.
        - Start with a question about an underlying *Population* -> collect a *Sample* -> use the *Sample* to calculate *Statistics* to estimate a *Parameter* of interest and make inferences about the *Population*; generally, the *Sample* should be representative of the *Population.*
      - *Variable* - a characteristic or measurement determined for each object in the *Population*
        - e.g. Student debt, GPA, major of 100 U.S. college students
        - 2 types:
          - *Numerical/Quantitative Variable* - e.g. amount of debt, GPA.; 2 types:
            - *Discrete* - countable; can't be divided in half forever; e.g. number of children in a household
            - *Continuous* - not countable; can take on any value within a given interval; can be divided in half forever; e.g. Age
          - *Categorical Variable* - e.g. major; observations can only fall into a particular category; 2 types:
            - *Nominal* - categories have no natural ordering; e.g. college major; if there is *no* more or less of the outcome when you move across categories.
            - *Ordinal* - categories have a natural ordering; e.g. how satisfied a student is with their school library; if there *is* more or less of the outcome when moving across categories.
    - Reason for using samples: too costly/time-consuming to get data on the entire population

- Need a *Representative Sample* to perform valid statistical inference
  - To obtain: use *Random Sampling*; e.g. Simple Random Sampling, Systematic Random Sampling, Stratified Sampling, Clustered Sampling
  - A random sample is representative of the population, in theory. In practice, it may not be
  - Inappropriate sampling can lead to an unrepresentative sample and incorrect conclusions
  - There is an infinite number of *samples* (and *statistics*) that can be drawn from the same *Population* (to be discussed later).