Notes

Week 1

Module 1 Week 1B: Descriptive Statistics

Descriptive Statistics

- **Descriptive Statistics**:
  - Describe the characteristics of a dataset
  - Can involve a single variable; e.g. average student debt.
  - Can involve multiple variables; e.g. correlation between average student debt and the GPA involves two variables.
  - Can be numerical, graphical, tabular.
  - Help a researcher know the data to know how to analyze it; otherwise, hard to navigate large sets of numbers, but key numerical measures/graphs are more insightful.
- **Bar Chart**:
  - Helpful for categorical data.
  - Categories typically on the x-axis; outcomes of interest – on the y-axis.
  - Sometimes y-axis measures the relative frequency of the categories.
- **Histogram**:
  - Adjoins intervals (boxes or bins) along the x-axis.
    - Tradeoff between clarity and precision: more bars is higher precision, but potentially worse clarity and vice versa.
  - Relative frequency (number of times a value occurs out of all data points) of the values within each interval on the y-axis.
  - If $f =$ frequency a values occurs; $n =$ sample size; then $\frac{f}{n} =$ relative frequency.
- **Time Series Graph (Line Graph)**:
  - Helpful to see how a variable behaves over time.
  - Time is typically along the x-axis.
  - Variable of interest is typically along the y-axis.
  - Can also do a split between multiple categories; e.g. unemployment rate across time between Alabama vs. Maryland.
  - Saint-Louis Federal Reserve's FRED Data Base has many data good for time-series analysis.
- **Measures of Location** - used to get a sense of a typical value for a variable:
  - **Percentiles**:
    - Find value i which corresponds to the k'th percentile: $i = \frac{k}{100}(n+1)$
    - E.g. 50th percentile value for 101 observations: $i = \frac{50}{100}(101+1) = 51$
      - I.e. the median corresponds to the 51st value of the variable.
    - Note: values must be ordered from smallest to largest.
  - **Quartiles** - separate the data into quarters.
    - Q1 - 25th percentile; Q2 - 50th percentile; Q3 - 75th percentile.
    - **Interquartile Range (IQR)** $= Q3 - Q1$.
    - Use quartiles to identify outliers in the data; rule of thumb:
      - Potential outlier if $(1.5 * IQR)$ below Q1 or above Q3.
  - **Boxplots** - graphical device to show location and spread of data.
    - The "box" represents the IQR, the middle 50% of the data.
    - The middle line represents the median (i.e. 50th percentile).
    - Can also show potential outliers.
- **Measures of Central Tendency** - helpful to determine what values are central or typical for data.
  - **Median** - the 50th percentile; if even number of data points, is the mean between two middle values (when values are sorted in ascending order).
  - **Mode** - the most frequently occurring value.
  - **(Arithmetic) Mean** - most common measure of central tendency.

    - $\bar{x}$ - typically denotes *Sample Mean*
    - Σ - summation.
    - **Sample Mean**:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

    - Note: other means exist, e.g. *Geometric Mean*, *Harmonic Mean*.
    - The *Mean* can be a misleading measure of central tendency when data are skewed.
      - If there is a strong skew to the data, **Median** is a preferred measure of the central tendency over the **Mean**; not as strongly influenced by the tails.
      - **Positive Skew -** i.e. right skew, the data has a long right tail.
      - **Negative Skew -** i.e. left skew, the data has a long left tail.
      - **Symmetric Shape -** no skew.
  - Relationship between *Mean* and *Median*:
    - If $\bar{x} > Med$ => positive skew.
    - If $\bar{x} < Med$ => negative skew.
- **Sample Variance -** the spread of the sample data.
  -

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

    - $(n-1)$: Bessel's correction; allows for a more reliable estimate of the variance
  - Measures how far observations typically fall from the mean; higher variance => more spread.
  - **Standard Deviation**: $s = \sqrt{s^2}$
    - Better units for interpretation than *Variance* because *Variance* is has a square in it.
    - For most variables, most values are within 2 *Standard Deviations* of the mean.
    - More precisely, **Chebyshev's Inequality**: $(1 - \frac{1}{k^2})$ of a variable's values (by proportion) must be within k *Standard Deviations* from the mean, for $k > 1$; regardless of distribution.
    - Difficult to compare variables measured in different units => **Standardize** variables to *Standard Deviation* units to compare:
      - Subtract the mean from each value and divide by *Standard Deviation*:
        - $Z = \frac{x - \bar{x}}{s}$
        - A *Standardized Variable* has a *Mean* of 0 and a *Standard Deviation* of 1.
        - Variable standardization will be used later in *Statistical Inference*.
- Important: distinguish between a *Population Parameter* (a quantity of interest) and a *Sample Statistic* (which we calculate and observe to do inference about a *Population Parameter*).
  - Common notation to denote the differences:
    - **Sample Mean**: $\bar{x}$
    - **Population Mean**: μ
    - **Sample Variance**: s^2
    - **Population Variance**: σ^2
  - Remember: *Random Sampling* makes our *Statistics* (e.g. $\bar{x}$) *Random Variables*
  - **Standard Error -** a *Statistic*'s *Standard Deviation*.
    - A *Statistic*'s variation can be calculated.