MIS2502: Data Analytics

Descriptive Statistics
Descriptive Statistics:

• Tools for summarizing, organizing, and simplifying data
• What data tells us about the population
  – Measures of Central Tendency
  – Measures of Dispersion
  – Tables & Graphs
Sample vs. Population

Population

Sample
Central tendency

• Mean (average)
  \[ \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \]

  Notation:
  \[ n = \text{the number of values} \]
  \[ x_1, x_2, \ldots, x_n = \text{the values} \]

• Median
  – The “middle” of a sorted list of numbers.

• Mode
  – The value that appears most often.
Dispersion

• Range
  – i.e. max - min

• Variance

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

• Standard deviation

\[ s = \sqrt{\frac{1}{n - 1} \sum_{i=1}^{n} (x_i - \bar{x})^2} \]
Histogram

A histogram is a graphical representation of the distribution of data.
Skewness is a measure of the asymmetry of the distribution.

Normal Distribution

- Symmetric
- Bell-shaped
Hypothesis testing

**Hypothesis Testing:** A technique for using data to validate/invalidate a claim about a population.

**Example 1:**
Population mean (Is the average delivery time of 30 mins really true?)

- Null Hypothesis: \( \mu = 30 \) (The average delivery time is 30 mins)
- Alternative Hypothesis: \( \mu \neq 30 \) (The average delivery time is different from 30 mins)

**Example 2:**
The difference in two population means (Is it true that the average income is the same in the neighborhood A versus neighborhood B?)

- Null Hypothesis: \( \mu_A = \mu_B \) (The average income is the same in the neighborhood A versus neighborhood B)
- Alternative Hypothesis: \( \mu_A \neq \mu_B \) (The average income is different in the neighborhood A versus neighborhood B)
Hypothesis testing uses *p-values* to weigh the strength of the evidence (what the data are telling you about the population).

The p-value is a number between 0 and 1.

- **A small \( p \)-value (typically \( \leq 0.05 \))** indicates strong evidence against the null hypothesis, so you **reject the null hypothesis**.

- **A large \( p \)-value (> 0.05)** indicates weak evidence against the null hypothesis, so you **fail to reject the null hypothesis**.