Topic 1: The Basics of Statistics

Definitions and Terms

- **Population:** An entire group of people (includes everyone in the entire population).
- **Sample:** The part of a population that we have data for and that we can use to examine the population (a subset of the population).

- When describing a population, we describe characteristics as **parameters.**
- **When describing a sample (as a subset of a population), we use the term statistic.**

- **The number of people in a population is denoted as N, whereas the number of people in a sample is denoted as n.**

Measures of Central Location

- The **mean** measures the average of the population/sample.

- The **median** measures the middle value of ordered (in increasing order) observations.
  - This is more sensitive to outliers than the mean;
  - However not as good as the mean if the data is skewed;
  - When the mean and median are equal, the data is symmetrical.

- The **range** is a measure of variability in the data.
  - This tells us about the spread of the data.
  - The maximum - minimum = range.

- The **variance** is a better measure of variability than the range.
  - Measures the average square distance from the mean (denoted by sigma squared).
  - Variance^2 = Sigma $((x-\mu)^2/N)$
  - Denoted as s^2 for samples.

- The **standard deviation** measures the spread on the original units of the data (unlike the variance which measures it in squared units).
  - This value is just the square root of the variance.

Coefficient of Variation

- The coefficient of variation is a relative measurement of variability with no true units.
- The value is equal to s / x bar.
- This basically compares the value of the standard deviation to the sample mean.

Percentile Location

- Creates percentile ranks like the median in relation to the entire set of data.
- Difference between the 75th and 25th percentile is called the interquartile range, where:
  - 25th percentile is the lower quartile range;
  - 50th percentile is the median;
  - 75th percentile is the upper quartile range.
- Percentile Rank = (Total Number of Observations - Rank Number) / Total Number of Observations

Means of Association

- **Covariance**: numerical variation of the correlation between two variables.
- It is just like calculating the variance twice.
- sigma $_{xy}$ = (Sigma (Point of X - Mean of X) • (Point of Y - Mean of Y)) / N
- If we are applying this to a sample, use the denominator n - 1 instead of N.
- This value tells us whether there is a positive or negative **linear** association.
- If the value is zero, there is **no linear association**, but not necessarily no relationship at all.
- However, this value has a units issue, and therefore if we are comparing two variables with different units, we would rather use the **correlation coefficient**.

- **Correlation Coefficient:** A standardised, unit free measure of association.
- p = sigma $_{xy}$ / (sigma $_{x}$ • sigma $_{y}$)
- If applied to a sample instead of a population, p is denoted as r, and sigma is denoted as s.
- A value of 1 tells us that there is a perfect positive linear relationship.
- A value of -1 tells us that there is a perfect negative linear relationship.
- This may be used to test the least squares method: a mathematical model to estimate the value of one variable given another.